Redbooks
ibm.com/redbooks

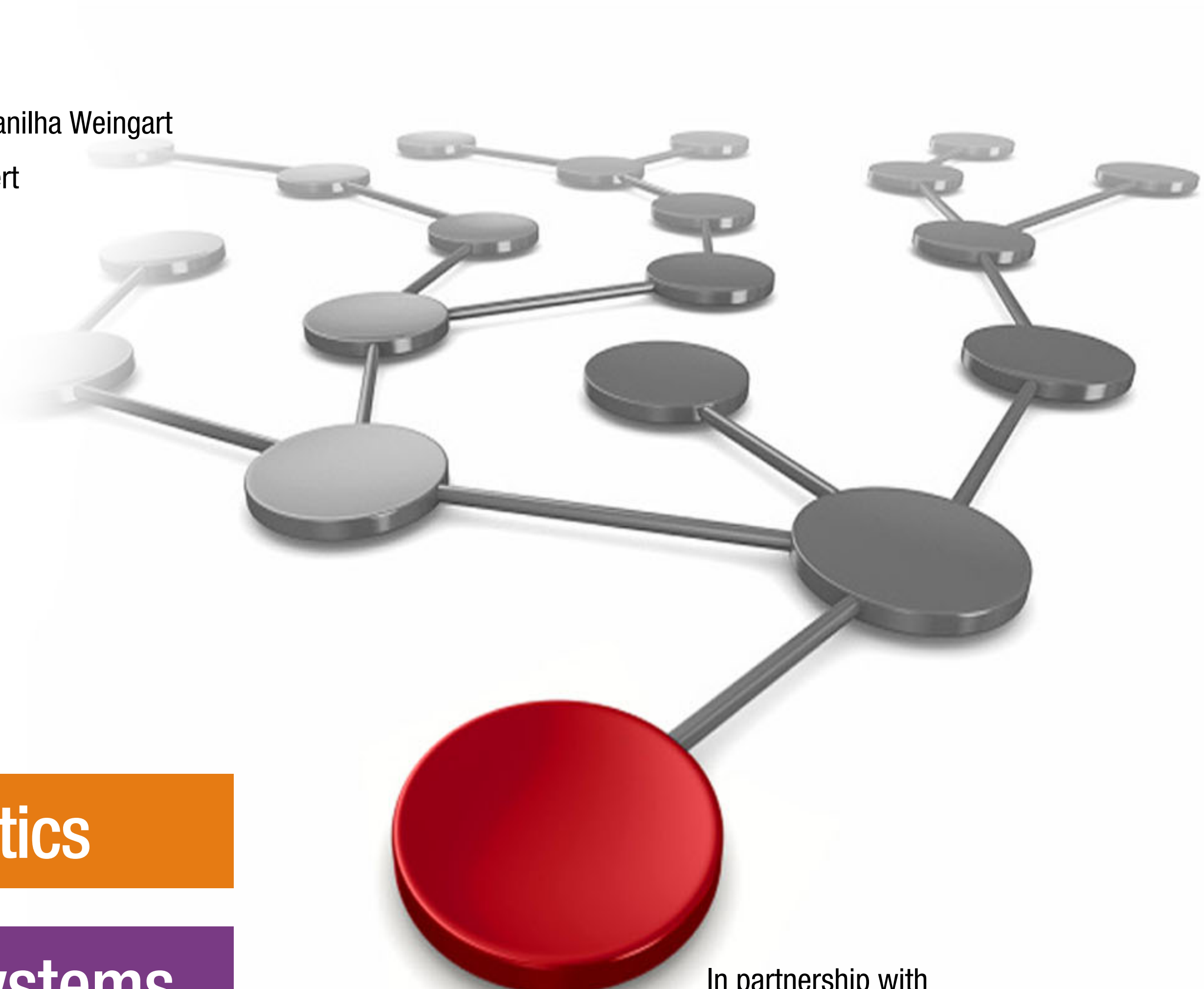
SAP HANA Data Management and Performance on IBM Power Systems

Dino Quintero

Adriana Melges Quintanilha Weingart

Marc-Stephan Tauchert

Faisal Siddique



Analytics

Power Systems

In partnership with
IBM **Academy of Technology**

IBM

Redpaper

IBM Redbooks

SAP HANA Data Management and Performance on IBM Power Systems

July 2021

REDP-5570-01

Note: Before using this information and the product it supports, read the information in “Notices” on page v.

Second Edition (July 2021)

This edition applies to:

- SAP HANA V2.0 SPS4 R44.
- SUSE Linux Enterprise Server V15 SP1.
- Red Hat Enterprise Linux V8.2.
- Hardware Management Console V9.1.940.0.

Contents

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

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

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US

INTERNATIONAL BUSINESS MACHINES CORPORATION 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. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

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

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

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

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.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

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 actual people or business enterprises 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. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and ibm.com are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the web at “Copyright and trademark information” at http://www.ibm.com/legal/copytrade.shtml

The following terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

AIX®	IBM Cloud®	Redbooks®
Db2®	POWER®	Redbooks (logo) ®
FlashCopy®	POWER9™	System z®
IBM®	PowerVM®	XIV®

The following terms are trademarks of other companies:

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

The registered trademark Linux® is used pursuant to a sublicense from the Linux Foundation, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis.

Red Hat, are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries.

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

Preface

This IBM® Redpaper® publication provides information and concepts about how to use SAP HANA and IBM Power Systems features to manage data and performance efficiently.

The target audience of this paper includes architects, IT specialists, and systems administrators who deploy SAP HANA and manage data and SAP system performance.

Authors

This paper was produced in close collaboration with the IBM SAP International Competence Center (ISICC) in Walldorf, SAP Headquarters in Germany and IBM Redbooks.

Dino Quintero is a Power Systems Technical Specialist with Garage for Systems. He has 25 years of experience with IBM Power Systems technologies and solutions. Dino shares his technical computing passion and expertise by leading teams that are developing technical content in the areas of enterprise continuous availability, enterprise systems management, high-performance computing (HPC), cloud computing, artificial intelligence (including machine and deep learning), and cognitive solutions. He also is a Certified Open Group Distinguished IT Specialist. Dino holds a Master of Computing Information Systems degree and a Bachelor of Science degree in Computer Science from Marist College.

Adriana Melges Quintanilha Weingart is an IBM Thought Leader and The Open Group Distinguished Technical Specialist certified, who is working as an Infrastructure Architect for SAP solutions on IBM Cloud®, reviewing exceptions and proposing viable alternatives to the solution architects and customers as part of the Boarding Solutions team. With more than 22 years of experience in IT/SAP, and as an IBM employee for 16 years, she also supported Global, LA, and Brazilian customers in Banking and Consumer Products industries as an SAP and Middleware Subject Matter Expert, working closely with customers, Business Partners, and other IBM teams. Adriana is a member of IBM Academy of Technology and the IBM Technology Leadership Council in Brazil. She has authored other IBM Redbooks® publications and participates as a speaker in IBM and non-IBM technical conferences.

Marc-Stephan Tauchert is a System Specialist at IBM Germany. In his over 26 years of experience in designing and implementing SAP Solutions on IBM Power Systems, he has covered various topics and roles. As a Technical Sales Consultant and SAP Solution Architect, he supports customer situations about 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. He is member of the IBM HANA on Power Systems Development Team in Sankt-Leon-Rot.

Faisal Siddique is an IBM Power System Lab Services Technical Specialist since 2011. Faisal specialties are IBM Power Systems, Linux on Power, SAP HANA on Power, and Spectrum Scale, including the IBM® Elastic Storage Server. Faisal has implemented major SAP HANA projects in MEA and MEP.

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

Wade Wallace
IBM Redbooks, Poughkeepsie Center

Katharina Probst, Walter Orb, Tanja Scheller
IBM Germany

Thanks to the authors of the previous editions of this paper:

Damon Bull
Vinicius Cosmo Cardoso
Cleiton Freire
Eric Kass

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an IBM Redbooks residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our papers to be as helpful as possible. Send us your comments about this paper or other IBM Redbooks publications in one of the following ways:

- Use the online **Contact us** review Redbooks form found at:

  ibm.com/redbooks

- Send your comments in an email to:

  redbooks@us.ibm.com

- Mail your comments to:

  IBM Corporation, IBM Redbooks
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- Look for us on LinkedIn:

  `http://www.linkedin.com/groups?home=&gid=2130806`

- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  `https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm`

- Stay current on recent Redbooks publications with RSS Feeds:

  `http://www.redbooks.ibm.com/rss.html`

1

Introduction

This chapter introduces the features of SAP HANA on IBM Power Systems that help manage data and performance efficiently.

This chapter includes the following topics:

1.1 Approaching SAP HANA on IBM Power Systems

In October 2019, the contributing authors to this publication gathered in a small office at the IBM SAP International Competence Center (ISICC) to define the content that is suitable for the audience of this publication. This audience includes system architects who are seeking guidance and best practices for designing and implementing SAP HANA on IBM Power Systems.

The authors of this publication are architects and engineers who are implementing SAP HANA systems for years, and often are asked to provide their insights about designing systems. The inquiries are diverse, the conditions are varied, and the answers are individual. However, in designing this book, the team attempted to anticipate the questions that are most likely asked.

The authors intend that this book be used in a task-based fashion to find answers to questions, such as the following examples:

- Which hardware do you choose?
- How do you plan for backups?
- How do you minimize start time?
- How do you migrate data to SAP HANA?
- How do you connect to remote systems?
- How much hardware do you need (memory, CPU, network adapters, and storage)?
- How do you reduce the memory footprint?
- What high availability (HA) options exist?
- Where do you get support?
- How do you answer other questions?

Consider this paper as a starting guide to learn what questions to ask and to answer the last question first. For any remaining questions, contact the ISICC.

The authors understand the goal of every SAP HANA installation is to be as resilient, available, inexpensive, fast, simple, extensible, and manageable as possible within constraints, even if some of the goals are contradictory.

This publication is unique in its intention to exist as a living document. SAP HANA technology and Power Systems technology that supports SAP HANA change so quickly that any static publication is out-of-date only months after distribution. The development teams intend to keep this book as up-to-date as possible.

So, where does the process to define the requirements of an SAP HANA system begin? Before any sizing is performed, your business goals must be established that come from the changing IT industry.

The typical motivation for moving forward to an SAP HANA-based system is the path toward digital transformation, which requires systems to perform real-time digitalization processing. For example, the requirements for processing pervasive user applications that use SAP Fiori and real-time analytics has a different processing footprint (Open Data Protocol [OData]-based) than a classical form-based process before output and the process after input into SAP GUI applications.

IBM bases its sizing on tables that are established by the benchmark team, who are highly proficient in SAP performance analysis. Conceptually, calculating SAP HANA sizing from a classical AnyDB system begins by determining the SAP Application Performance Standard (SAPS), which is a standard measure of workload capacity, of your present system.

Then, that value is used as reference for the IBM sizing tables to determine the hardware requirements for an Power Systems server to run SAP HANA with an equivalent SAPS performance.

After you size the memory, you find that the number of hardware permutations that fits your requirements is overwhelming. To narrow your choices, you decide whether to use a scale-up or scale-out infrastructure. As a best practice, use a *scale-up* infrastructure because although some automation is available to help select which data can be spanned across multiple nodes, you still need some manual management when you implement a *scale-out* structure.

When you use a scale-up infrastructure, consider that different hardware implementations have different performance degradation characteristics when scaling-up memory usage. As systems become larger, the memory to CPU architecture plays an important role in the following situations:

- How distant the memory is from the CPU (*affinity*).
- How proficiently the system can manage cache to memory consistency.
- How well the CPU performs virtual address translation of large working sets of memory.

Power Systems are designed to scale up by using enterprise-grade, high-bandwidth memory interconnects to the CPU, and relatively flat non-uniform memory access (NUMA). Scaling up with Power Systems servers is a best practice to support large memory footprint SAP HANA installations (for more information, see 2.2.5, “SAP HANA on IBM Power Systems and SAP HANA Data Tiering solutions” on page 20). The SAP HANA on Power Systems scaling benefits are complemented by the advantage of consolidating multiple SAP HANA, Linux, IBM AIX®, and IBM i workloads on the same machine.

Note: Acquisition costs and licensing become complicated when running mixed workloads. Ask your customer representative to contact the ISICC to have them help provide an optimal cost configuration. Contacting the ISICC is a good idea in general because we want to know what you are planning, and it is our job to help make the offering as attractive as possible to fit your needs.

Customers that have SAP incidence support for SAP on IBM products continue to enjoy that support with SAP HANA. SAP HANA on Power Systems support channels are intricately integrated into SAP development and support.

The people who are supporting SAP on AIX, SAP with IBM Db2®, or SAP on IBM i, and System z® are members of the same team that support SAP HANA on Power Systems.

By migrating to SAP HANA but remaining with IBM, support teams were not changed. For questions regarding anything SAP, open an SAP support ticket with an IBM coordinated support channel, such as BC-OP-AIX, BC-OP-AS4, BC-OP-LNX-IBM, BC-OP-S390; and for issues regarding interaction with other databases (DBs), BC-DB-DB2, BC-DB-DB4, and BC-DB-DB6.

1.1.1 Memory footprint

In contrast to the disk-storage-based DB engines that are used by classical SAP Enterprise Resource Planning (ERP), a memory footprint with memory-based DB systems is a continuous issue. A memory footprint is not a static calculation because the calculation takes time. If your archiving rate cannot keep up with your data generation rate, your memory footprint tomorrow is greater than it is today.

Data growth is an issue that continuously requires attention because it is likely that your application suite changes. Therefore, the characteristics of your data generation also likely change.

An archiving strategy is your primary application-level control for data growth, but some applications, such as SAP Business Warehouse (BW), support near-line-storage as an alternative. For data that is not archived for various reasons, SAP HANA supports a selection of hardware technologies for offloading data that is not used frequently (warm data or cold data) to places other than expensive RAM with expensive SAP licensing.

The available technologies to alleviate large data and data growth are SAP HANA Native Storage Extension (NSE), which is a method of persisting specific tables on disk and loading data into caches as necessary (it is similar in function to classical DB engines), extension nodes (slower scale-out nodes), and SAP HANA Dynamic Tiering (SAP HANA to SAP HANA near-line storage).

Note: Some of the options that are offered by SAP HANA are not available for all applications. SAP S/4HANA features restrictions other than SAP BW, and both SAP S/4HANA and SAP BW includes more restrictions than SAP HANA native applications.

SAP provides reports to run on your original system to provide data placement assistance, and the results are typically good suggestions. You must be prepared to distribute your data differently if prototypes demonstrate other configurations are necessary.

For more information about planning SAP archiving and managing various technologies for data that is accessed at different frequencies, see Chapter 2, “SAP HANA data growth management” on page 7.

1.1.2 Start times

The SAP HANA processing model implies that data is in memory for immediate access. The cost of accessing data from storage is met when the SAP HANA system starts, during which all data is loaded into RAM. For example, if your storage is connected by 4-multipath 16 Gbps lines, loading 7 TB requires 1 hour. BM and Intel provide solutions to minimize start times by using persistent memory in which data survives SAP HANA restarts in specific situations.

Although the hardware and operating system methods of retaining persistent data in memory varies, in SAP HANA persistent memory is referenced as memory mapped files. The operating system or hardware provides a memory-based file system (such as a RAM disk) that is “seen” as a file system that is mounted in a path.

Instructed by a configuration parameter, SAP HANA uses as much persistent storage as possible and reverts to regular memory for the remainder. SAP HANA can be instructed to attempt to store everything in persistent memory, or each file can be altered to prefer persistent or non-persistent RAM.

Even with hardware support for a quick start, consider that an HA solution alleviates the need to wait for SAP HANA to restart because SAP HANA is always available. With an HA solution in place, the preferred method for scheduled maintenance is to move the data to the backup site.

1.1.3 Backup

SAP HANA provides an interface for backup tools that is called *Backint*. SAP publishes a list of certified third-party tools that conform to that standard. If you plan on using methods, such as storage-based snapshot or flash copies, quiesce the SAP HANA system before taking a storage-based snapshot. A quiesce state is necessary to apply subsequent logs (the record of transactions that occur since the time of the quiesce) to an image that is restored from the IBM FlashCopy® copy.

1.1.4 High availability

SAP HANA System Replication (HSR) is the underlying foundation of DB HA. The method is based on *logical replication*, where individual row-level changes are recorded and transferred to the backup site (referred to as *log shipping*). Logical replication is the opposite of *storage-based replication*, where pages on disk are duplicated without considering the logical data structure of the change.

SAP HSR accomplishes the task of transferring changes between hosts by transferring segments of the log (segments of the record of the changes to the DB that are used for transactional recovery). Changes that are received by backup hosts are sent to the local DB in a fashion that is similar to when undergoing forward recovery.

1.1.5 Exchanging hardware

Unfortunately, exchanging hardware is a problem that you encounter every few years. If your system manages to fulfill all the prerequisites for Live Partition Mobility (LPM), which is a feature of Power Systems hardware that moves running partitions between hosts, exchanging hardware by using LPM is the least disruptive to a running SAP system. However, the prerequisites are difficult to master. Many specialized hardware features, which are not virtualized and must be assigned to a physical partition, make LPM impossible.

The SAP standard method of exchanging hardware (for example, any hardware with the same endianness), is by using SAP HSR. For more information, see SAP Note 1984882.

> **Note:** Some links for more information throughout the publication t require to have an S-ID to access them. For more information about SAP user and authorization, see this SAP web page.

1.1.6 Remote database connectivity

SAP provides a wealth of data connectivity options. Some methods, such as Remote Function Call (RFC) and Database Multiconnect, are provided at the layer of the SAP system. Other methods, such as Smart Data Access (SDA) and Smart Data Integration (SDI), are integrated directly into SAP HANA. Extract-transform-load (ETL) methods include SAP Landscape Transformation (SLT) and the SAP Data Migration Option (DMO) for upgrades.

SDA and SDI are not specific to SAP HANA to IBM i for the connectivity function. The IBM i case is a prime example of the use of a generic Open Database Connectivity (ODBC) adapter for SDA, and the generic Camel JDBC adapter for SDI.

1.1.7 Conclusion

Facilitating designs of superior SAP HANA systems is the goal of this publication.

Although your productive, development, and test systems include different reliability, availability, serviceability (RAS); connectivity; and security requirements, the aspects to consider are universal.

Where tradeoffs must be made between cost, RAS, and complexity are decisions that are unique to your situation, the intent of this paper and the service and support you receive from the ISICC and the IBM SAP development team is to help optimize your decisions so that you feel comfortable and confident with the final architecture of your design.

2

SAP HANA data growth management

This chapter describes the challenges of data growth in systems and how it can affect the operations of an SAP HANA database.

This chapter introduces the data temperature concept, which is used as a guide to decouple data types that are based on their criticality. Data temperature is helpful for companies that are deciding when to move their data to different but still accessible data tiers.

Different SAP data tiering solutions that are supported on IBM Power Systems servers are described in this chapter. The purpose of this chapter is to help you decide on the most suitable solution among the different available solutions.

This chapter includes the following topics:

2.1 The challenge of data growth management and the data temperature concept

Companies collect more information about their business to control their day-to-day operations in real time. The more information that is collected, the more IT resources are used over time, which increases the costs for organizations because of the need for data scaling.

In an SAP HANA database, main memory and disk areas are used, which increases the total cost of ownership (TCO) and affects performance over time.

Before scaling up or scaling out the SAP HANA database, think about options for decoupling your data location by defining what data always must be in memory and available with the highest performance for applications and users. Also, consider what data is less frequently accessed so that it available from a lower performance data tier with no effect on the business operations.

You can define which data is accessed infrequently so that it can be available to users in a reasonable and cheaper performing storage tier. This concept is called the *data temperature*.

Figure 2-1 shows how data temperature is classified.

HOT	Data read and written very frequently for mission-critical business tasks and available for real-time processing and real-time analytics.
WARM	Data less frequently accessed with reduced performance so that it can be stored in a lower cost tier with reduced SLA.
COLD	Very infrequent access, historical data, typically read-only. Data stored on low cost storage tiers, like disk only, but still accessible at anytime.

Figure 2-1 Data temperature concept

The use of data tiering options for your SAP HANA database includes the following benefits:

- Reduce data volume and growth in the hot store and SAP HANA memory.
- Avoid performance issues on SAP HANA databases because too much data must be loaded into the main memory.
- Avoid needing to scale up or scale out over time.
- Ensure lower TCO.

SAP offers the following data tiering solutions that are supported on SAP HANA on IBM Power Systems:

- Near-Line Storage (NLS) (cold data)
- SAP HANA Extension Node (warm data)
- SAP HANA Native Storage Extension (NSE) (warm data)
- SAP HANA Dynamic Tiering (warm data)

2.2 SAP HANA data tiering options for SAP HANA databases

This section describes the different data tiering options for SAP HANA databases.

2.2.1 Near-Line Storage with SAP IQ

With NLS, you can move your SAP Business Warehouse (BW) data that is classified as cold out of the primary SAP HANA database and store it in SAP IQ.

SAP IQ is a column-based with petabyte scale relational database software system that is used for business intelligence, data warehousing, and data marts. Produced by Sybase Inc. (which is now an SAP company), its primary function is to analyze large amounts of data in a low-cost, high availability (HA) environment.

SAP developed NLS to use with SAP NetWeaver BW and SAP BW/4HANA *only*.

Prerequisites for Near-Line Storage

The following SAP NetWeaver BW versions and support package levels are required:

- SAP NetWeaver BW 7.30 SPS >= 09
- SAP NetWeaver BW 7.31 SPS >= 07
- SAP NetWeaver BW 7.4 SPS >= 00
- SAP NetWeaver BW 7.5 SPS >= 00

The required SAP BW/4HANA versions and support package level is SAP BW/4HANA 1.0 SPS >= 00.

For more information about the minimum release level for SAP BW, see SAP Note 1796393.

The NLS implementation includes the following components:

- SAP NetWeaver BW on SAP HANA database or SAP BW/4HANA
- SAP IQ database
- SAP IQ Libraries on SAP Application Server side for connection to SAP IQ
- (Optional) SAP Smart Data Access (SDA) on SAP HANA Server side for connection to SAP IQ

Implementing SDA to access the NLS data is optional. SAP HANA SDA optimizes running queries by moving processing as much as possible to the database that is connected through SDA. The SQL queries the work in SAP HANA on virtual tables. The SAP HANA Query Processor optimizes the queries and runs the relevant parts in the connected database; then, it returns the result to SAP HANA and completes the operation.

The use of an NLS solution with SAP HANA SDA is supported as of SAP NetWeaver BW 7.40 SP8 or higher or SAP BW/4 HANA 1.0 or higher.

> **Note:** To use the SDA solution, the SAP BW application team must configure the SAP BW objects.

For more information about the use of SAP HANA SDA for NLS and its performance benefits, see the following SAP Notes:

- SAP Note 2165650
- SAP Note 2156717
- SAP Note 2100962

NLS implementation architecture

The architecture of NLS implementation with SAP NetWeaver BW on SAP HANA and SAP BW/4HANA without SDA is shown in Figure 2-2.

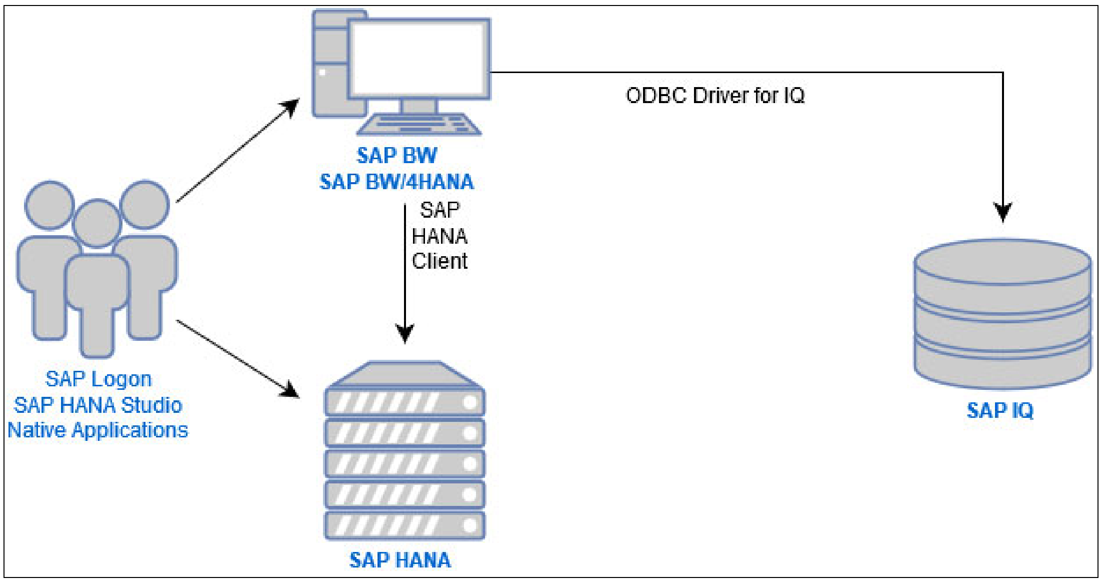


Figure 2-2 NLS architecture without Smart Data Access

The architecture of NLS implementation with SAP NetWeaver BW on SAP HANA and SAP BW/4HANA with SDA is shown in Figure 2-3.

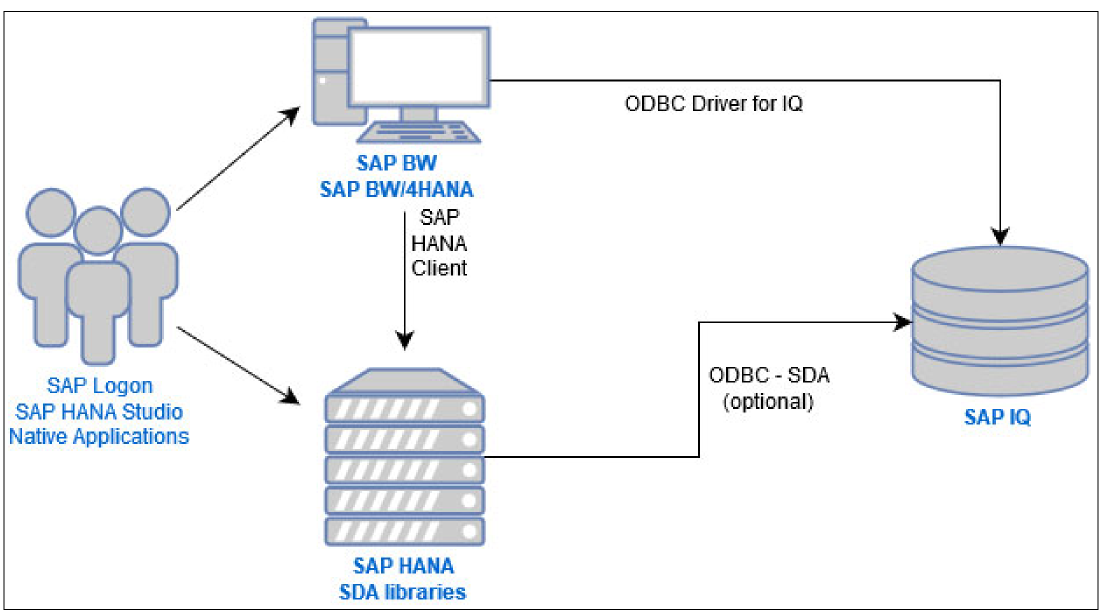


Figure 2-3 NLS architecture with Smart Data Access

> **Hint:** SAP IQ is supported on Power Systems (Big Endian only); therefore, it is possible to reserve a logical partition (LPAR) for implementing it and the use of a virtual LAN (VLAN) to produce the lowest network latency possible between SAP IQ and SAP HANA.

For the SDA implementation, SAP developed and provided the packages that are supported on Power Systems servers.

For more information about the implementation of NLS for SAP BW, see SAP Note 2780668.

2.2.2 SAP HANA Extension Node

The SAP HANA Extension Node is based on the SAP HANA scale-out feature. It stores the warm data of SAP BW in a separate node from the SAP HANA main nodes.

Sizing for the SAP HANA Extension Node

In addition to use of the memory area for storing static objects, such as tables (for example, the data footprint), SAP HANA also uses memory for dynamic objects (for example, objects for the SAP HANA run time). The SAP recommendation for SAP HANA memory sizing is to reserve as much memory for dynamic objects as for static objects in the column store.

> **Note:** As a best practice, the SAP recommendation for memory sizing is:
>
> `RAMdynamic = RAMstatic`

For example, if you use a footprint of 500 GB, your SAP HANA database memory size must be at least 1 TB.

The SAP HANA Extension Node can operate twice as much data with the same amount of memory and fewer cores. For example, if you expect to have up to 1 TB of footprint in your Extension Node, you can have up to 250 GB of memory for the Extension Node.

Prerequisites for the SAP HANA Extension Node

For SAP BW on HANA, you need SAP HANA 1.0 SPS12 or later and SAP BW 7.4 SP12 or later. However, SAP BW 7.50 SP05 or SAP BW/4HANA is recommended. For SAP HANA native applications, you need SAP HANA 2.0 SPS3 or later.

The ideal use case for Extension Node is SAP BW on SAP HANA or SAP BW/4HANA because the SAP BW application controls the data distribution, partitioning, and access paths. For SAP HANA native applications, all data categorization and distribution must be handled manually.

SAP HANA Extension Node Deployment

A possible SAP HANA scale-out with Extension Node implementation is shown in Figure 2-4.

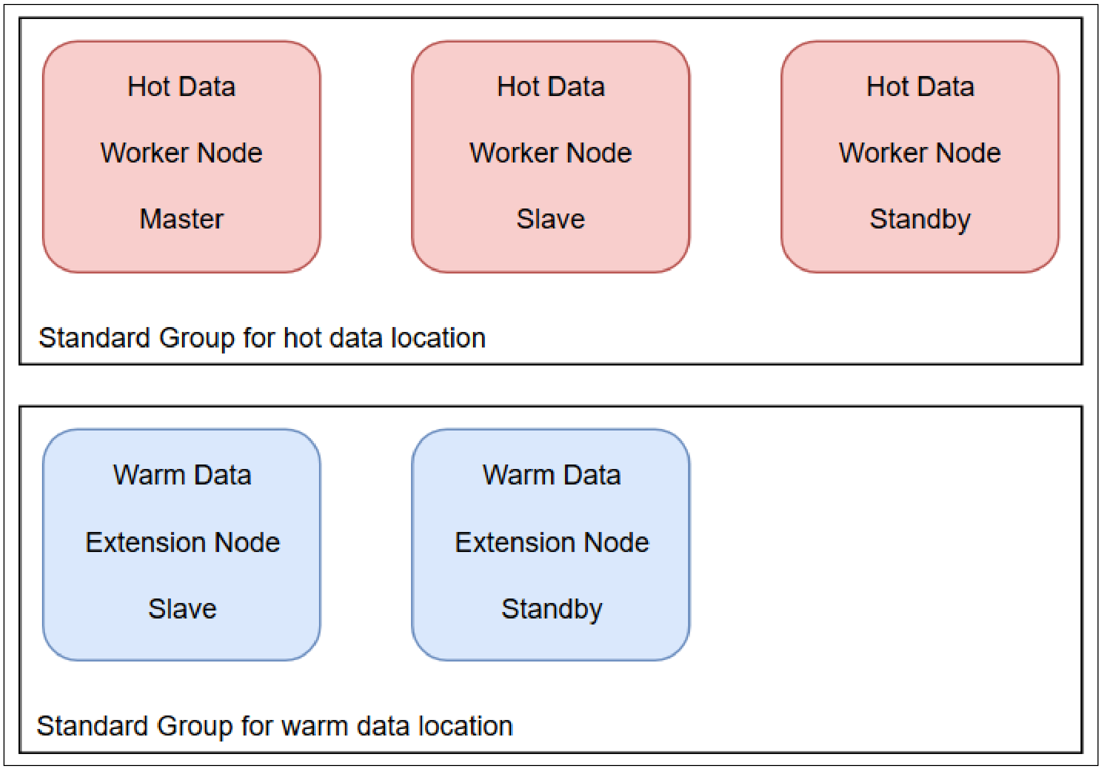


Figure 2-4 SAP HANA Extension Node Deployment option

SAP HANA Extension Node is built into the SAP HANA platform and supported on Power Systems.

For more information about the implementation of Extension Node, see the following SAP Notes:

- SAP Note 2486706
- SAP Note 2643763

2.2.3 SAP HANA Native Storage Extension

Starting with SAP HANA 2.0 SPS4, SAP developed a new native data tiering solution for warm data that is called *Native Storage Extension* (NSE). NSE can process in-memory stored data for performance of critical operations. It has a buffer-cache managed page to process less frequently accessed data.

To activate NSE, you must configure your warm data-related tables, or partitions of columns to be page-loadable, by running `SQL DDL` commands.

After the table, partition, or column is configured to use NSE, it is no longer loaded into memory. However, it is readable by using the buffer cache. The performance for accessing data on that table is slower than accessing it in-memory.

Total cost of ownership reduction and performance advantages

If you use NSE and make 50% of your data page-loadable, 50% of memory is freed from the SAP HANA main memory. Therefore, persistent area and buffer cache is used instead. This situation can reduce TCO because most of it is related to the amount of memory (DRAM) that is needed in the server.

Also, depending on the amount of data that is moved to NSE, SAP HANA can start faster during scheduled and unexpected maintenance.

Supportability

NSE supports SAP HANA native applications for SAP S/4HANA and SAP Suite on SAP HANA (SoH).

> **Note:** SAP recommends the use of NSE with SAP S/4HANA or SoH with Data Aging only. If Data Aging is used in SAP S/4HANA or SoH with SAP HANA 2.0 SPS4, NSE is used for storing the historical partitions of tables in aging objects. To learn more about this use case for SAP S/4HANA, see SAP Note 2816823.

The NSE Advisor

The NSE Advisor is bundled with NSE. Use the Advisor to learn what tables, partitions, or columns are suitable to be converted to page-loadable (to save the memory space) or column-loadable (to improve the performance) within the recommendations view result.

Based on the workload on the table, partition, or column over time, the NSE Advisor identifies the frequently accessed and rarely accessed objects so that the system administrators can decide which objects can be moved to NSE.

Figure 2-5 shows the NSE architecture perspective.

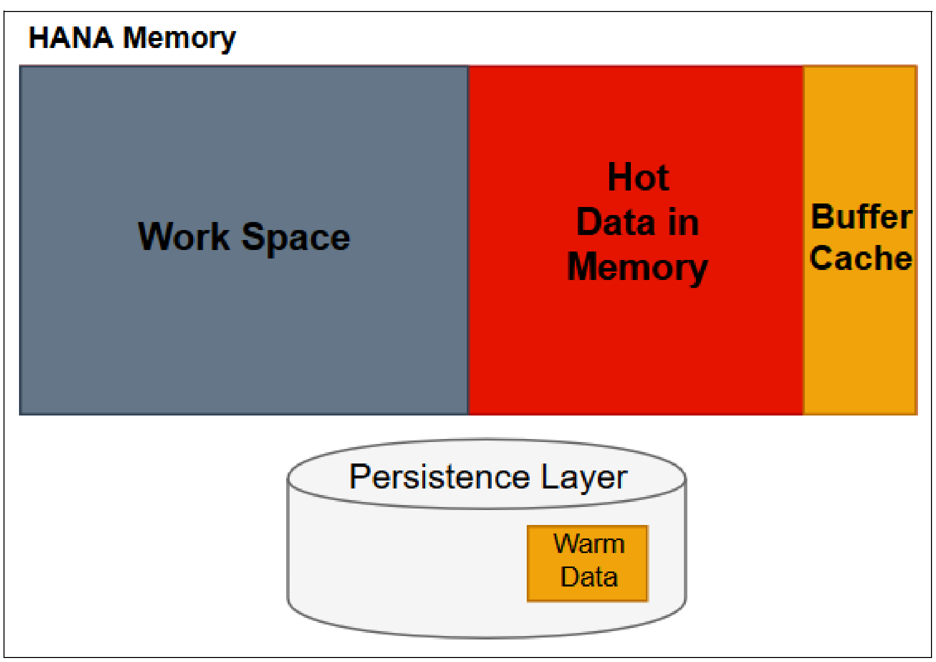


Figure 2-5 NSE architecture

NSE is included with SAP HANA 2.0 SPS4, and is supported on Power Systems. For more information about the use of NSE, see SAP Note 2799997.

2.2.4 SAP HANA Dynamic Tiering

Starting in SAP HANA 1.0 SPS09, SAP introduced *SAP HANA Dynamic Tiering*, which is another option for storing warm data on disk. The data is stored in a columnar format to dedicate more space in memory for hot data, which decreases the TCO of your SAP HANA database.

SAP HANA Dynamic Tiering is not included in the standard installation package; therefore, you must download an extra component and install it on SAP HANA.

In SAP HANA Dynamic Tiering, you can create two types of warm tables: extended and multistore. The *extended table* type is disk-based; therefore, all data is stored in disk. The *multistore table* type is an SAP HANA partitioned table with some partitions in memory and some partitions on disk.

The distribution of the data among the in-memory store tables, extended tables, and multistore tables is shown in Figure 2-6.

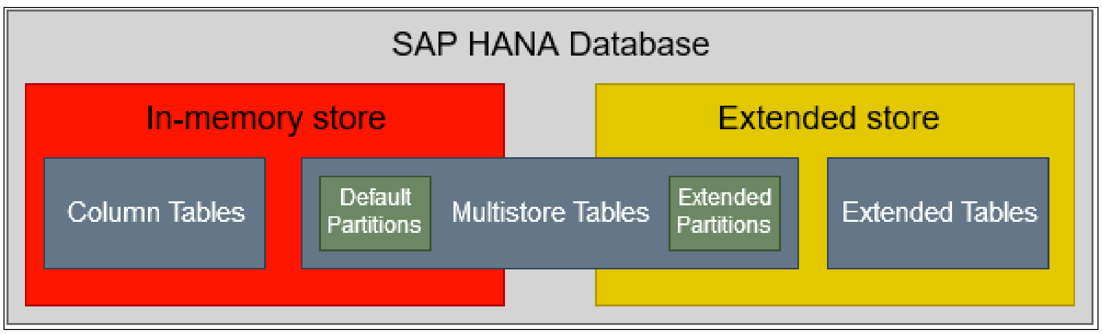


Figure 2-6 SAP HANA Dynamic Tiering: Distribution of tables

SAP HANA Dynamic Tiering can be installed in the same server that is hosting SAP HANA or in a separate dedicated host server. You can also install a second SAP HANA Dynamic Tiering host as a standby host for HA purposes.

The operating system process for the SAP HANA Dynamic Tiering host is **hdbesserver**, and the service name is **esserver**.

SAP HANA Dynamic Tiering supports only low-tenant isolation. Any attempt to provision the SAP HANA Dynamic Tiering service (**esserver**) to a tenant database with high-level isolation fails. After the SAP HANA Dynamic Tiering implementation, the SAP HANA Dynamic Tiering service stops working if you attempt to configure the tenant database with high isolation.

SAP HANA Dynamic Tiering configuration properties

The `esserver.ini` file stores the extended storage configuration properties. With these parameters, you can configure the SAP HANA Dynamic Tiering for memory usage, concurrent connections, concurrent queries, and so on.

Table 2-1 lists important parameters of SAP HANA Dynamic Tiering.

Table 2-1 SAP HANA Dynamic Tiering configuration properties

Startup Section Properties		
Parameter name	**Description**	**Default value**
`catalog_cache`	Amount of memory that is initially reserved for caching the dynamic tiering catalog.	32000000 bytes
`checkpoint_interval`	Maximum interval between checkpoints.	60 minutes
`es_log_threshold_size`	Size (in megabytes) of the log file threshold for point-in-time recovery in dynamic tiering.	Value is based on db space size. For databases less than 1TB, the threshold is 10% of database pace size or 2 GB, whichever is greater. For larger databases, the value is 1% of database space size.
`delta_memory_mb`	Amount of memory that is available to store a delta-enabled extended table.	2048 MB
`heap_memory_mb`	Amount of heap memory that dynamic tiering can use. A value of `0` or empty means no limit on heap memory.	1024 MB
`load_memory_mb`	Maximum amount of memory the extended storage can request from the OS for temporary use.	2048 MB
`main_cache_mb`	Amount of memory to be used for caching dynamic tiering database objects.	1024 MB
`max_concurrent_con¬nections`	Maximum number of concurrent connections that the dynamic tiering service accepts.	50
`max_concurrent_queries`	Maximum number of concurrent queries that are allowed by the server.	32
`num_partition_buffer_cache`	Number of main and temp buffer cache partitions. Must be a power of 2; otherwise, the value is rounded to the nearest power of 2 - 64.	None. Value is determined at runtime based on the number of CPUs, and is not user visible.
`num_threads`	Maximum number of threads that are used for dynamic tiering.	600
`temporary_cache_mb`	Amount of memory to be used as cache for temporary objects during dynamic tiering operations.	256

The dynamic tiering service must be restarted for changes to the database section properties to take effect, as listed in Table 2-2.

Table 2-2 Restart service for these SAP HANA Dynamic Tiering configuration properties to take effect

Database Section Properties		
Parameter name	**Description**	**Default value**
es_log_backup_interval	Specifies the time interval (in minutes) after which the eslog (log for the dynamic tiering extended store; these files are copied into the backup directory and represent the active log) is backed up. This property is set at the database level. Reducing the log backup time interval to less than the time it takes to perform a log backup has no effect because log backups do not run in parallel. Subsequent log backups are triggered after only the current log backup finishes. For example, if the current log backup takes 10 minutes to finish, setting `es_log_backup_interval=5` does not trigger the next log backup until the current backup finishes. For more information, see Log Backups under **System Administration → Backup and Recovery for Dynamic Tiering → Creating Backups** in the SAP HANA Dynamic Tiering: Administration Guide.	15 (minutes)
es_log_threshold_size	Specifies the minimum available size (in megabytes) of the partition (not the size of eslog) on which the eslog volume is mounted and allowed to grow. If the size of the partition falls under the specified size, dynamic tiering returns a warning or error and prompts you to free up or add space to the file system. This property is set at the SYSTEM level.	The default value is based on database space size. For databases less than 1 TB, the threshold is 10% of database space size or 2 GB, whichever is greater. For larger databases, the value is 1% of database space size.

The dynamic tiering service must be restarted for changes to the trace section properties to take effect, as listed in Table 2-3. These properties control the size and number of the `esserver` and `esserver_console` trace files that are in the HANA trace directory.

Table 2-3 Restart service for these SAP HANA Dynamic Tiering configuration properties to take effect

Trace Section Properties		
Parameter name	**Description**	**Default value**
maxfiles	Specifies the number of archives of the old message log that is maintained by the server. Applies only if maxfilesize is not 0.	10
maxfilesize	Limits the maximum size of the message log. The value is set in bytes.	10000000

The dynamic tiering service must be restarted for changes to the Zrlog section properties to take effect, as listed in Table 2-4. These properties control request-level logging for dynamic tiering and typically are used for diagnostic purposes only. To enable request-level logging, the `statement_type` property must be set to a value other than the default (NONE).

Table 2-4 Restart service for these SAP HANA Dynamic Tiering configuration properties to take effect

Zrlog Section Properties		
Parameter name	**Description**	**Default value**
filesize_limit	Create a log file and rename the original log file when the original log file reaches specified size.	0
maxfiles	Specify number of request log file copies to retain. Takes effect only if `filesize_limit` is also specified.	0
statement_type	Enable request logging of operations. Separate multiple values with a comma (,) or plus sign (+).	NONE
tracefile	Redirect request logging information to a file separate from the regular log file.	`trace/es_requestlog_$HOST_${PORT}_${COUNT:3}.log`

For a complete list of the parameters, see *SAP HANA Dynamic Tiering: Administration Guide*.

Deployment options for SAP HANA Dynamic Tiering

You can deploy SAP HANA Dynamic Tiering in a dedicated host, and in the same host of SAP HANA. An SAP HANA Dynamic Tiering service can coexist in the same SAP HANA host for only one SAP HANA tenant. Therefore, if you need SAP HANA Dynamic Tiering for any additional tenants, the **`esserver`** service must be in a different host.

Same host deployment

In this host deployment, you have just one SAP HANA host with the tenant (**`indexserver`**) and SAP HANA Dynamic Tiering (**`esserver`**) services in the same SAP HANA host. Starting with SAP HANA Dynamic Tiering 2.0, the same host deployment is supported for production environments. Figure 2-7 shows the configuration scenario.

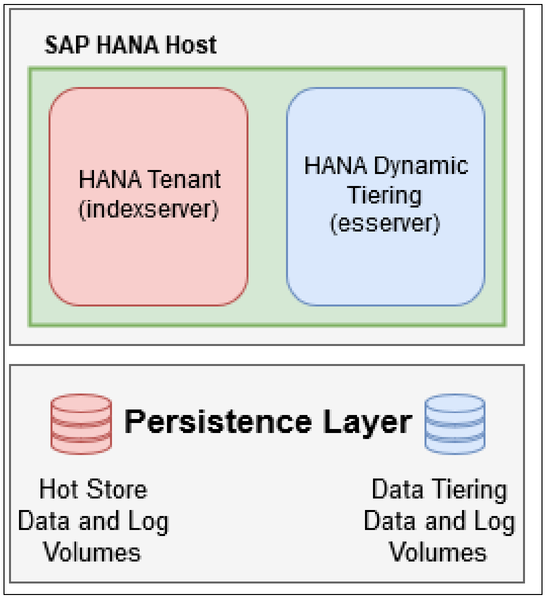


Figure 2-7 SAP HANA Dynamic Tiering: Same host deployment

> **Note:** Some host deployments are designed for small, nonproduction environments, but are supported in production environments.

Dedicated host deployment

In a dedicated host deployment, the SAP HANA tenant service and SAP HANA Dynamic Tiering service are in different SAP HANA hosts, and the SAP HANA Dynamic Tiering service of host B is associated with the SAP HANA tenant service of host A (see Figure 2-8).

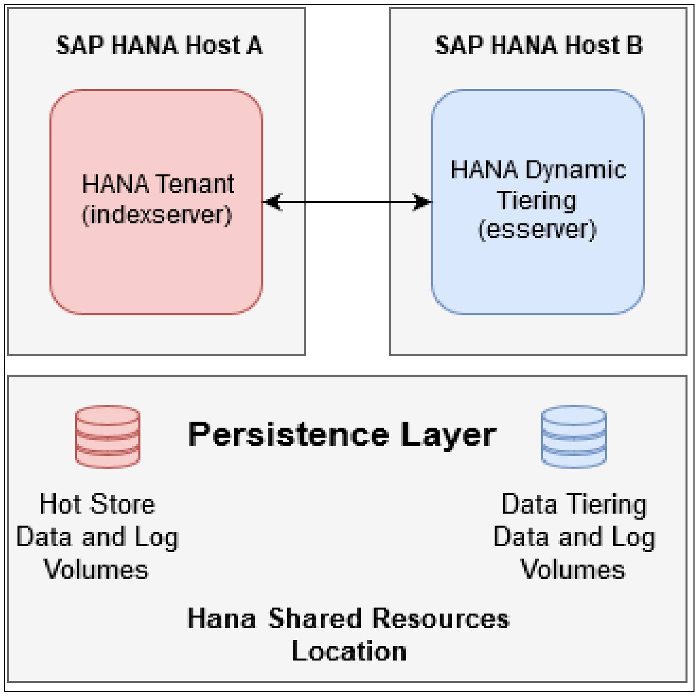


Figure 2-8 SAP HANA Dynamic Tiering: Dedicated host deployment

This deployment is where the organizations can use all Power Systems benefits with the flexibility of LPAR support, with low network latency.

More than one SAP HANA Dynamic Tiering server deployment

In this scenario, two SAP HANA tenant databases are running in a same SAP HANA system. In this case, if SAP HANA Dynamic Tiering is needed for the two tenant databases, the SAP HANA Dynamic Tiering service for the second tenant database must be in an extra host because the SAP HANA tenant service (`indexserver`) and the SAP HANA Dynamic Tiering service (`esserver`) can coexist on the same host for only one tenant. Figure 2-9 shows the configuration for this scenario.

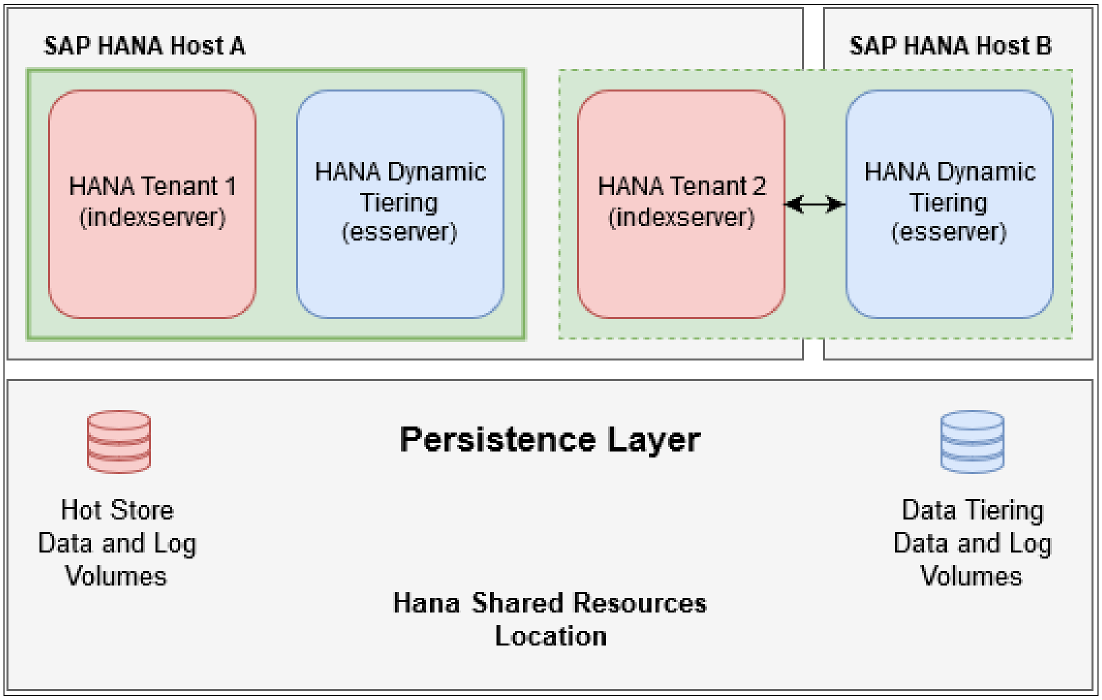


Figure 2-9 More than one SAP HANA Dynamic Tiering server deployment

2.2.5 SAP HANA on IBM Power Systems and SAP HANA Data Tiering solutions

With Power Systems, a scale-up or scale-out SAP HANA database environment can be implemented.

With support for multiple LPARs, organizations can consolidate multiple SAP HANA instances or SAP HANA nodes of the same instance (multi-host) on a single Power Systems server by using its simplicity of management with a low-latency network.

By using IBM PowerVM® (the Power Systems hypervisor), you can virtualize up to 16 production SAP HANA instances on the LPARs of a single Power Systems server (IBM Power Systems E950 or IBM Power Systems E980 server). It is also possible to move memory and CPUs among the LPARs with flexible granularity (for more information, see SAP Note 2230704). This Note is being updated by SAP.

PowerVM allows more granular scaling and dynamically changing allocation of system resources. You can avoid adding hardware that can cause higher energy, cooling, and management needs.

The SAP HANA on Power Systems solution runs the same SUSE or Red Hat Linux distributions as x86 servers, with the flexibility, scalability, resiliency, and performance advantages of Power Systems servers that help the client to accomplish the following tasks:

- Accelerate SAP HANA deployments by using the flexibility of built-in virtualization, which allows faster provisioning of SAP HANA instances and allocating capacity with granularity as little as 0.01 cores and 1 GB.
- Minimize infrastructure and simplify management by using the following capabilities:
  - Virtualization scalability of up to 24 TB in scale-up mode.
  - The ability to deploy up to 16 SAP HANA modules in a single server.
  - Shared processor pools that optimize CPU cycles across SAP HANA virtual machines (VMs) in a server.

Power Systems is the best solution for implementing SAP HANA scale-up and scale-out modes and the data tiering solutions that are shown in this section. For more information, see SAP HANA Server Solutions with Power Systems.

2.2.6 SAP HANA Dynamic Tiering: Hands-on demonstration

This section demonstrates the implementation of SAP HANA Dynamic Tiering.

Before you start

This section describes a high-level, step-by-step installation of SAP HANA Dynamic Tiering on SAP HANA 2.0 on Power Systems (Little Endian). It also includes a demonstration about how to create an extended storage and multistore table for manipulating data on it.

You use the Console Interface (CI) to perform an SAP HANA Data Tiering on a Same Host Deployment; that is, on the same server where the SAP HANA is installed with no additional SAP HANA nodes.

Note: Some data manipulation is demonstrated by using SAP HANA Studio.

For more information and detailed procedures, see *SAP HANA Dynamic Tiering: Master Guide*, *SAP HANA Dynamic Tiering: Installation and Update Guide*, and *SAP HANA Dynamic Tiering: Administration Guide* at SAP HANA Dynamic Tiering.

Preparing the installation

This section describes how to prepare for the installation.

Checking the hardware and operating system requirements

According to the SAP HANA Dynamic Tiering Installation Guide, SAP HANA Dynamic Tiering is available for:

- Intel -based hardware platforms
- Power Systems servers

Note: Power Systems environments require the suitable IBM XL C/C++ redistributable libraries. Download and install the suitable runtime environment for the latest updates from the supported IBM C and C++ Compilers page at the IBM Support Portal. Install the libraries on the SAP HANA and SAP HANA Dynamic Tiering hosts. These libraries are not required for an Intel -based hardware platform environment.

Checking SAP HANA compatibility with the SAP HANA Dynamic Tiering version

For a matrix of compatible SAP HANA and SAP HANA Dynamic Tiering versions, see SAP Note 2636634 .

Downloading the SAP HANA Dynamic Tiering package

To download the SAP HANA Dynamic Tiering package, go to the SAP Software Download Center and select **Support Packages and Patches** → **By Alphabetical Index (A-Z)** → **H** → **SAP HANA DYNAMIC TIERING** → **SAP HANA DYNAMIC TIERING 2.0** → **COMPRISED SOFTWARE COMPONENT VERSIONS** → **SAP HANA DYNAMIC TIERING 2.0**. At the top of the download page, click **LINUX ON POWER LE 64 BIT**.

Now, you can download the SAP HANA Dynamic Tiering revision that is compatible with the SAP HANA 2.0 SP level you have in place or are installing.

Extracting the SAP HANA Dynamic Tiering package

After you download and copy the package to the SAP HANA host, extract it by running the following command:

```
SAPCAR -manifest SIGNATURE.SMF -xvf DYNTIERING20004_3-70002283.SAR
```

Installing SAP HANA Dynamic Tiering (same host deployment)

To install SAP HANA Dynamic Tiering by using the CI, complete the following steps:

1. Log in to the SAP HANA host by using the root ID and change to the resident `hdblcm` directory within the `<sid>` folder; for example, `/hana/shared/TST/hdblcm`, where `/hana/shared/` is the shared resources location for the system and TST is the HANA `<sid>`.
2. Start the SAP HANA platform lifecycle management tool by running the command that is shown in Example 2-1, where `<full-path-option>` refers to the path where SAP HANA Dynamic Tiering installation package was extracted.

Example 2-1 SAP HANA Dynamic Tiering installation command for starting the installation

```
./hdblcm --component_dirs=/<full_path_option>
```

3. At the `Choose an action` prompt, select **Install** or **Update Additional Components**, as shown in Example 2-2.

Example 2-2 SAP HANA Dynamic Tiering installation: Choose an action window

```
SAP HANA Lifecycle Management - SAP HANA Database 2.00.048.03.1605873454
************************************************************************

Choose an action

  Index | Action                     | Description
  -----------------------------------------------------------------------------------
  1     | add_host_roles             | Add Host Roles
  2     | add_hosts                  | Add Hosts to the SAP HANA Database System
  3     | check_installation         | Check SAP HANA Database Installation
  4     | configure_internal_network | Configure Inter-Service Communication
  5     | configure_sld              | Configure System Landscape Directory Registration
  6     | extract_components         | Extract Components
  7     | print_component_list       | Print Component List
  8     | remove_host_roles          | Remove Host Roles
  9     | rename_system              | Rename the SAP HANA Database System
  10    | uninstall                  | Uninstall SAP HANA Database Components
```

```
  11    | unregister_system           | Unregister the SAP HANA Database System
  12    | update                      | Update the SAP HANA Database System
  13    | update_component_list       | Update Component List
  14    | update_components           | Install or Update Additional Components
  15    | update_host                 | Update the SAP HANA Database Instance Host
integration
  16    | exit                        | Exit (do nothing)

Enter selected action index [16]: 14
Scanning software locations...
Detected components: SAP HANA Dynamic Tiering (2.00.048.03.1605873454) in
/home/DynamicTearing/es
```

4. At the `Choose components to be installed or updated` prompt, select **Install SAP HANA Dynamic Tiering**, as shown in Example 2-3.

Example 2-3 SAP HANA Dynamic Tiering installation: Choose components to be installed or updated window

```
Choose components to be installed or updated:

  Index | Components | Description

-------------------------------------------------------------------------------
  1     | all        | All components
  2     | es         | Install SAP HANA Dynamic Tiering version
2.0.043.00.12711

Enter comma-separated list of the selected indices: 2
```

5. You are prompted to add another host. Enter `n`, as shown in Example 2-4.

Example 2-4 SAP HANA Dynamic Tiering installation: Add hosts window

```
Verifying files...
Do you want to add hosts to the system? (y/n) [n]:
```

6. You are prompted to enter the System Database User Name and password. Enter `SYSTEM` and its password, as shown in Example 2-5.

Example 2-5 SAP HANA Dynamic Tiering installation: User and password window

```
Enter System Database User Name [SYSTEM]:
Enter System Database User (SYSTEM) Password:
```

7. You are prompted to add the paths for SAP HANA Dynamic Tiering data and log volume paths. In this case, the paths are `/hana/data/dtes/TST` for the data volumes and `/hana/log/dtes/TST` for the log volumes, as shown in Example 2-6.

Example 2-6 SAP HANA Dynamic Tiering: Installation

```
Enter Location of Dynamic Tiering Data Volumes [/hana/data_es/TST]:
/hana/data/dtes/TST
Enter Location of Dynamic Tiering Log Volumes [/hana/log_es/TST]:
/hana/log/dtes/TST
```

8. You are prompted to confirm all the added parameters. Enter `y`, as shown in Example 2-7.

Example 2-7 SAP HANA Dynamic Tiering installation: Summary window

```
Summary before execution:
=========================

SAP HANA Database
   Update Parameters
      Remote Execution: ssh
      Update Execution Mode: standard
      System Database User Name: SYSTEM
      Location of Dynamic Tiering Data Volumes: /hana/data/dtes/TST
      Location of Dynamic Tiering Log Volumes: /hana/log/dtes/TST
   Software Components
      SAP HANA Dynamic Tiering
         Install version 2.0.048.03.14465
         Location: /home/DynamicTearing/es
   Log File Locations
      Log directory:
/var/tmp/hdb_TST_hdblcm_update_components_2020-12-07_01.44.16
      Trace location: /var/tmp/hdblcm_2020-12-07_01.44.16_22005.trc

Do you want to continue? (y/n): y
```

During the installation process, you see messages that are similar to the examples that are shown in Example 2-8.

Example 2-8 SAP HANA Dynamic Tiering installation process

```
Installing components...
Installing SAP HANA Dynamic Tiering...
  Installing jre8...
  Installing shared...
  Installing lang...
  Installing conn_lm...
  Installing conn_add_lm...
  Installing odbc...
  Installing client_common...
  Installing server...
  Installing complete - log files written to /hana/shared/TST/es/log
  Restarting HANA System...
  Copying delivery unit HANA_TIERING
  Copying delivery unit HDC_TIERING
Updating Component List...
SAP HANA Database components updated
```

The installation log files are in the following path:

```
/var/tmp/hdb_TST_hdblcm_update_components_<date_and_time>/hdblcm.log
```

SAP HANA Dynamic Tiering postinstallation activities

This section describes the postinstallation steps after the SAP HANA Dynamic Tiering installation.

Adding the SAP HANA Dynamic Tiering role

To add the SAP HANA Dynamic Tiering role, complete the following steps:

1. While logged in to the SAP HANA host with root ID, go to the path `/hana/shared/TST/hdblcm`, run the SAP HANA Lifecycle Management tool, and select **Add Host Roles**, as shown in Example 2-9.

Example 2-9 SAP HANA Dynamic Tiering role: Running the SAP HANA Lifecycle Management tool

```
./hdblcm
SAP HANA Lifecycle Management - SAP HANA Database 2.00.048.03.1605873454
************************************************************************

Choose an action

  Index | Action                       | Description

---------------------------------------------------------------------------------------
---
  1     | add_host_roles               | Add Host Roles
  2     | add_hosts                    | Add Hosts to the SAP HANA Database System
  3     | check_installation           | Check SAP HANA Database Installation
  4     | configure_internal_network   | Configure Inter-Service Communication
  5     | configure_sld                | Configure System Landscape Directory Registration
  6     | extract_components           | Extract Components
  7     | print_component_list         | Print Component List
  8     | remove_host_roles            | Remove Host Roles
  9     | rename_system                | Rename the SAP HANA Database System
  10    | uninstall                    | Uninstall SAP HANA Database Components
  11    | unregister_system            | Unregister the SAP HANA Database System
  12    | update                       | Update the SAP HANA Database System
  13    | update_component_list        | Update Component List
  14    | update_components            | Install or Update Additional Components
  15    | update_host                  | Update the SAP HANA Database Instance Host
integration
  16    | exit                         | Exit (do nothing)

Enter selected action index [16]: 1
```

2. Select the SAP HANA host to be assigned the extra SAP HANA Dynamic Tiering role. In this case, only one host is available, as shown in Example 2-10.

Example 2-10 SAP HANA Dynamic Tiering role: Host selection for adding role

```
System Properties:
TST /hana/shared/TST HDB_ALONE
        HDB00
        version: 2.00.048.03.1605873454
        host:  linux-y743 (Database Worker (worker))
        edition: SAP HANA Database

Select hosts to which you would like to assign additional roles

  Index | System host | Roles
  ----------------------------------------------
```

```
  1     | linux-y743      | Database Worker (worker)

Enter comma-separated list of selected indices [1]:
```

3. In `Select additional host roles for host '<host>'`, select the host. In this case, only t one host is available. Add the `<sid>adm` ID password for it, as shown in Example 2-11.

Example 2-11 SAP HANA Dynamic Tiering role: Selection of host

```
Select additional host roles for host ' linux-y743'

  Index | Additional Role         | Role Description

-------------------------------------------------------------------------------
---
  1     | extended_storage_worker | Dynamic Tiering Worker
(extended_storage_worker)

Enter comma-separated list of additional roles for host ' linux-y743' [1]:
Enter System Administrator (TSTadm) Password:
```

4. You are prompted to confirm all added parameters. Confirm and enter `y`, as shown in Example 2-12.

Example 2-12 SAP HANA Dynamic Tiering role: Installation summary window

```
Summary before execution:
=========================

Add Host Roles
   Add Host Roles Parameters
      Do not start hosts with modified roles: No
      Remote Execution: ssh
      Auto Initialize Services: Yes
      Do not Modify '/etc/sudoers' File: No
   Additional Host Roles
       linux-y743
         Current Role(s): worker
         Additional Role(s): extended_storage_worker
   Log File Locations

Log directory: /var/tmp/hdb_TST_hdblcm_add_host_roles_2020-12-07_01.56.10
      Trace location: /var/tmp/hdblcm_2020-12-07_01.56.10_28958.trc

Do you want to continue? (y/n): y
```

At the end of the process, you see a summary of the installation, as shown in Example 2-13.

Example 2-13 SAP HANA Dynamic Tiering role: Installation process

```
Assigning Additional Roles to the Local Host...
  Adding role 'extended_storage_worker' on local host 'linux-y743'...
  Performing esaddhost...
    esaddhost.sh: Configuring SAP ES...
  Stopping instance...
```

```
    Stopping 8 processes on host 'linux-y743' (extended_storage_worker, 
worker):
      Stopping on 'linux-y743' (extended_storage_worker, worker): hdbdaemon, 
hdbcompileserver, hdbesserver, hdbindexserver (TST), hdbnameserver, 
hdbpreprocessor, hdbwebdispatcher, hdbxsengine (TST)
    All server processes stopped on host 'linux-y743' (extended_storage_worker, 
worker).
  Stopping sapstartsrv service...
  Starting service (sapstartsrv)...
  Starting instance TST (HDB00) on host 'linux-y743'...
    Starting 8 processes on host 'linux-y743' (extended_storage_worker, 
worker):
      Starting on 'linux-y743' (extended_storage_worker, worker): hdbdaemon, 
hdbcompileserver, hdbesserver, hdbindexserver (TST), hdbnameserver, 
hdbpreprocessor, hdbwebdispatcher, hdbxsengine (TST)
      Starting on 'linux-y743' (extended_storage_worker, worker): hdbdaemon, 
hdbesserver, hdbindexserver (TST), hdbwebdispatcher, hdbxsengine (TST)
      Starting on 'linux-y743' (extended_storage_worker, worker): hdbdaemon, 
hdbwebdispatcher, hdbxsengine (TST)
      Starting on 'linux-y743' (extended_storage_worker, worker): hdbdaemon, 
hdbwebdispatcher
    All server processes started on host 'linux-y743' (extended_storage_worker, 
worker).
Additional host roles successfully assigned
```

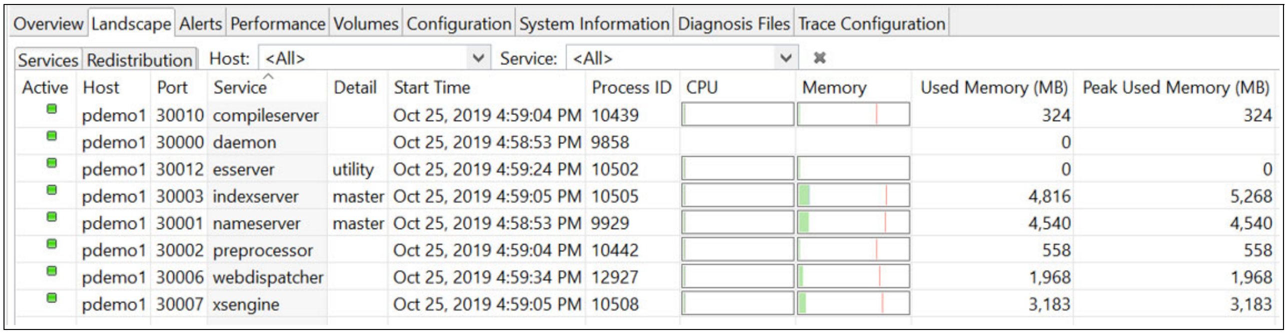
The installation log files are in the following path:

```
/var/tmp/hdb_TST_hdblcm_add_host_roles_<date_and_time>/hdblcm.log
```

5. Using SAP HANA Studio, log in to the tenant database as the SYSTEM ID. In the Overview tab, you see the SAP HANA Dynamic Tiering status as “Installed but not running yet” for that tenant, as shown in Figure 2-10.

Figure 2-10 SAP HANA Dynamic Tiering: Yellow status in SAP HANA Studio

6. Click the **Landscape** tab, and you see the SAP HANA Dynamic Tiering Server service **`esserver`**, as shown in Figure 2-11.



Overview | Landscape | Alerts | Performance | Volumes | Configuration | System Information | Diagnosis Files | Trace Configuration

Services | Redistribution | Host: <All> | Service: <All>

Active	Host	Port	Service	Detail	Start Time	Process ID	CPU	Memory	Used Memory (MB)	Peak Used Memory (MB)
	pdemo1	30010	compileserver		Oct 25, 2019 4:59:04 PM	10439			324	324
	pdemo1	30000	daemon		Oct 25, 2019 4:58:53 PM	9858			0	
	pdemo1	30012	esserver	utility	Oct 25, 2019 4:59:24 PM	10502			0	0
	pdemo1	30003	indexserver	master	Oct 25, 2019 4:59:05 PM	10505			4,816	5,268
	pdemo1	30001	nameserver	master	Oct 25, 2019 4:58:53 PM	9929			4,540	4,540
	pdemo1	30002	preprocessor		Oct 25, 2019 4:59:04 PM	10442			558	558
	pdemo1	30006	webdispatcher		Oct 25, 2019 4:59:34 PM	12927			1,968	1,968
	pdemo1	30007	xsengine		Oct 25, 2019 4:59:05 PM	10508			3,183	3,183

Figure 2-11 SAP HANA Dynamic Tiering: Service esserver in SAP HANA Studio

Creating extended storage

After the SAP HANA Dynamic Tiering installation and role additions are completed, extended storage must be created. The *extended storage* is a database space that is the file on disk where tables and partitions store the SAP HANA Dynamic Tiering.

To create the extended storage, you must have the system privileges EXTENDED STORAGE ADMIN.

> **Note:** The extended storage is created with SYSTEM ID now, but for all subsequent activities, you need another user ID with all necessary privileges, and it is the owner of the extended storage and multistore tables.

Complete the following steps:

1. On SAP HANA Studio, connect to the SAP HANA Tenant database to which the SAP HANA Dynamic Tiering was provisioned to with the SYSTEM ID.

   > **Note:** In this demonstration, a single tenant is the initial tenant, and the SAP HANA instance previously never contained more tenants. Therefore, the SAP HANA Dynamic Tiering service is automatically provisioned to the tenant database.

2. For this demonstration, create one extended storage with 1 GB of available allocated space. Right-click the tenant database and click **Open SQL Console** to open the SQL console. Then, run the command that is shown in Example 2-14.

   Example 2-14 SAP HANA Dynamic Tiering: Command for creating extended storage

   ```
   CREATE EXTENDED STORAGE AT ' linux-y743' SIZE 1000 MB;
   ```

Notice that `linux-y743` is the location (host) that is used in this demonstration. Adjust it to your host. The result is shown in Figure 2-12.

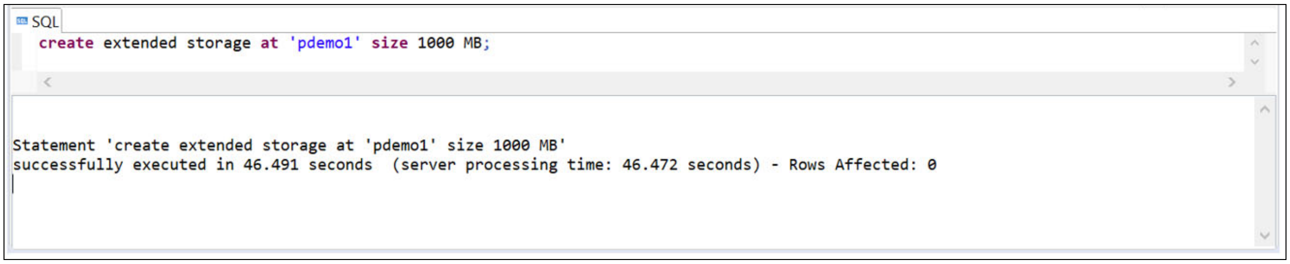


Figure 2-12   SAP HANA Dynamic Tiering: Result for extended storage creation command-line interface

3. Click the **Overview** tab in SAP HANA Studio, you see the status of SAP HANA Dynamic Tiering as “Running”, as shown in Figure 2-13.

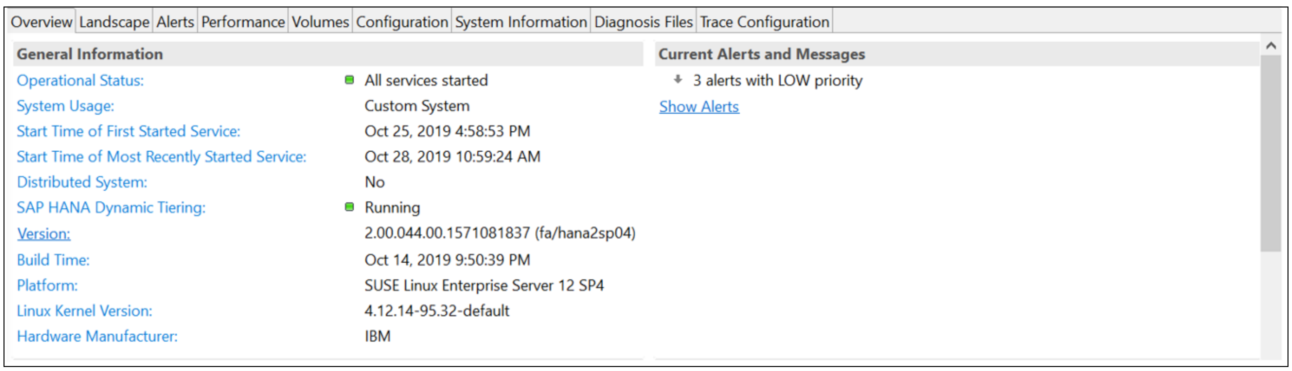


Figure 2-13   SAP HANA Dynamic Tiering: Running status in SAP HANA Studio

Your SAP HANA Dynamic Tiering is now ready for you to create an Extended Table or multistore table.

Creating a user ID for your SAP HANA Dynamic Tiering objects

You created your extended storage with the tenant SYSTEM ID. Now, to create your extended table and multistore tables, you need a new user ID that creates and owns these objects.

To create the user ID by using SAP HANA Studio, complete the following steps:

1. Click **Tenant** → **Security**, right-click **Users**, and then, click **New User**.
2. Define a name for the user and add the System Privileges CATALOG READ, EXTENDED STORAGE ADMIN, and IMPORT. You do not have to force the password change in the first log on if you prefer.

The parameters are shown in Figure 2-14.

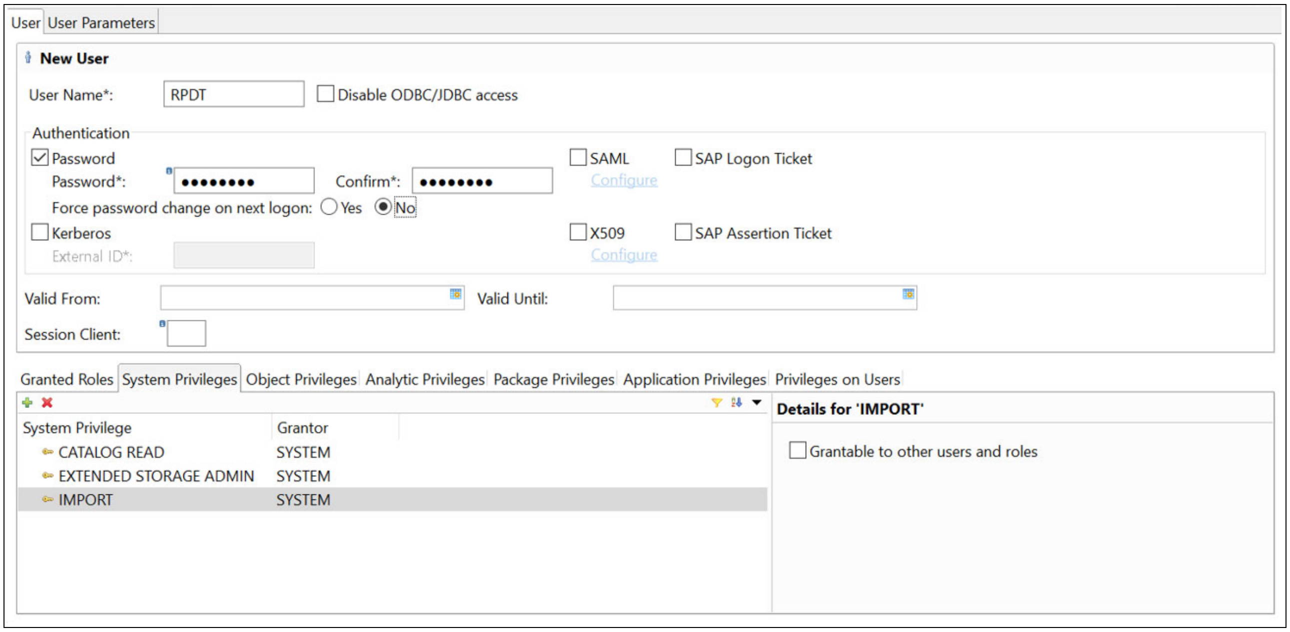


Figure 2-14 SAP HANA Dynamic Tiering: User ID creation properties

In this case, the user ID is DTUSER. Log in to the tenant with the user ID that was defined by you.

Creating an extended table

You can create an extended table in the same way you create any other column table in an SAP HANA database; the only difference is that you must add `USING EXTENDED STORAGE` at the end of the statement, as shown in Example 2-15.

Example 2-15 SAP HANA Dynamic Tiering: Extended table creation command-line interface

```
CREATE TABLE "DTUSER"."CUSTOMER_ES" (
C_CUSTKEY     integer          not null,
C_NAME        varchar(25)      not null,
C_ADDRESS     varchar(40)      not null,
C_PHONE       char(15)         not null,
primary key (C_CUSTKEY)
) USING EXTENDED STORAGE;
```

The same table was created in this demonstration, as shown in Figure 2-15.

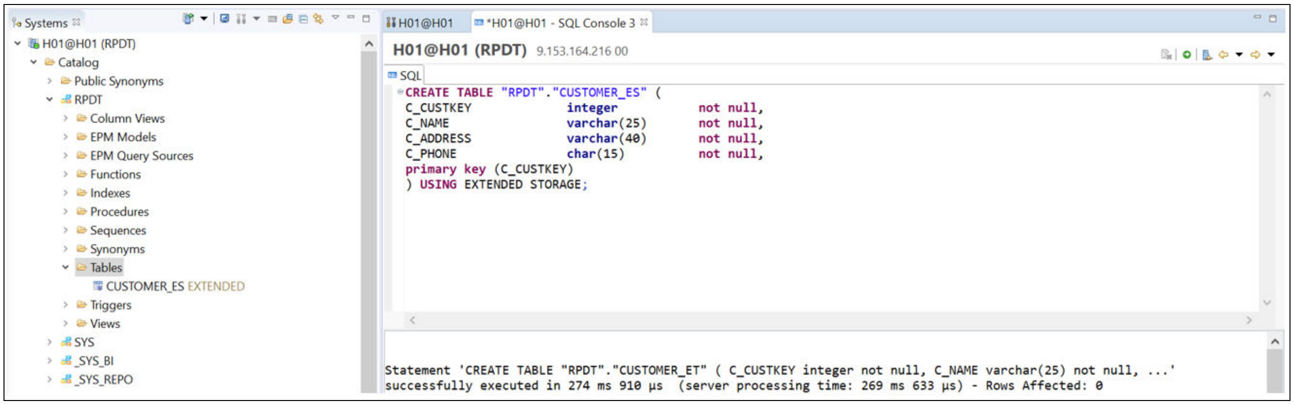


Figure 2-15 SAP HANA Dynamic Tiering: Extended table CUSTOMER_ES creation

In left pane of the window, the table is identified as an EXTENDED table in the SAP HANA Catalog for user DTUSER.

> **Important:** Foreign keys between two extended tables or between an extended table and an in-memory table are *not* supported.

To insert data, use the same syntax as a common in-memory column store table, as shown in Example 2-16 and in Figure 2-16.

Example 2-16 SAP HANA Dynamic Tiering: Insert data into table CUSTOMER_ES command-line interface

```
INSERT INTO "DTUSER"."CUSTOMER_ES"
(C_CUSTKEY, C_NAME, C_ADDRESS, C_PHONE)
VALUES
(1,'CUSTOMER 1','ADDRESS 1', 19999999999);
```

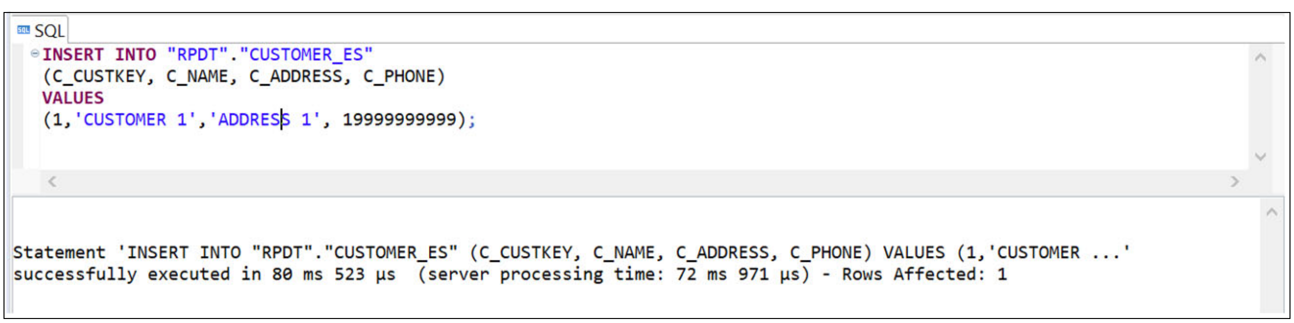


Figure 2-16 SAP HANA Dynamic Tiering: Insert data into table CUSTOMER_ES - SAP HANA Studio window

To read data from the table, the SQL syntax that you use is the same for reading data from any other table. Therefore, you can select data in the same way as any other in-memory column store table, as shown in Figure 2-17.

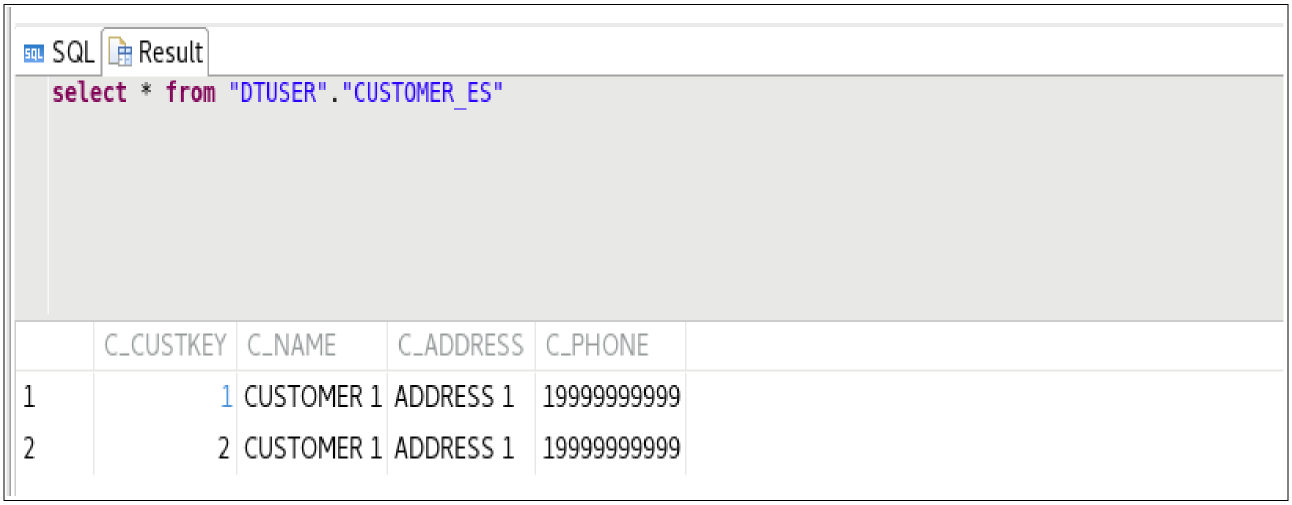


Figure 2-17 SAP HANA Dynamic Tiering: Contents of the extended table CUSTOMER_ES in SAP HANA Studio

Creating a multistore table

When creating a multistore table, you must split it in partitions and define which partitions are part of the default store (for example, the memory store) and which partitions are part of the extended storage.

Complete the following steps:

1. Create a table called SALES_ORDER and partition it by RANGE by using a date type field as the partition field. In the default store, the values are 2010/12/31 - 9999/12/31, and in the extended storage, the values are 1900/12/31 - 2010/12/31, as shown in Example 2-17.

   Example 2-17 SAP HANA Dynamic Tiering: Creating a multistore table command-line interface

   ```
   CREATE TABLE "DTUSER"."SALES_ORDER" (
   S_SALESOKEY            integer          not null,
   S_CUSTOMER             integer          not null,
   S_VALUE                decimal(15,2)    not null,
   S_DATE                 date             not null,
   primary key (S_SALESOKEY,S_DATE))
   PARTITION BY RANGE ("S_DATE")
   (
   USING DEFAULT STORAGE
   (PARTITION '2010-12-31' <= VALUES < '9999-12-31')
   USING EXTENDED STORAGE
   (PARTITION '1900-12-31' <= VALUES < '2010-12-31'));
   ```

   In SAP HANA Studio, you see that the multistore table symbol differs from the extended table symbol, as shown in Figure 2-18 on page 33.

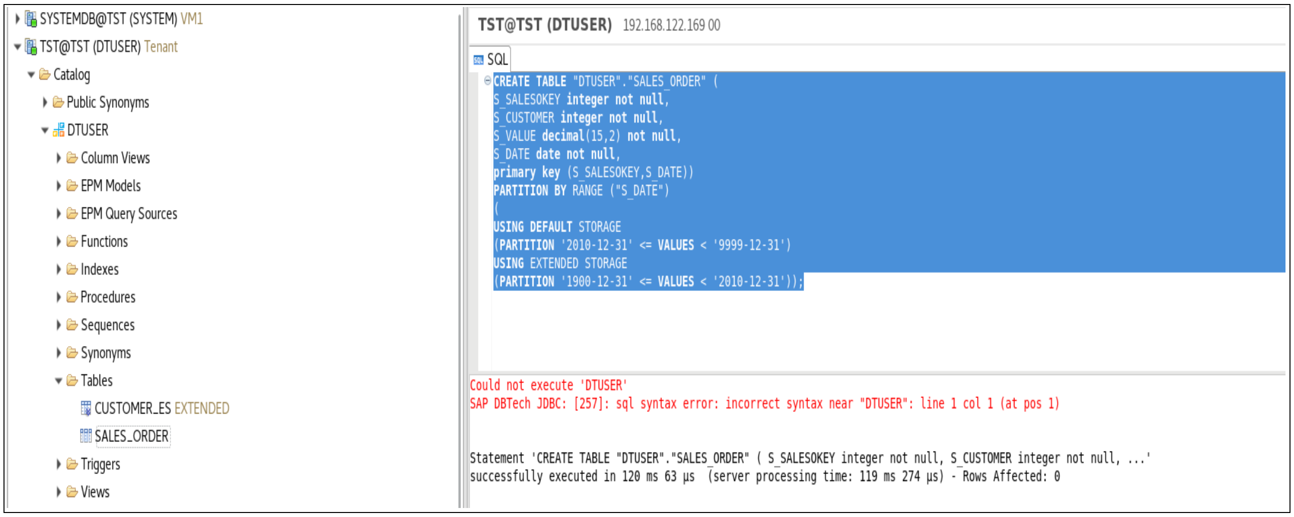


Figure 2-18 SAP HANA Dynamic Tiering: Multistore table SALES_ORDER creation

The Data Manipulation Language (DML) operations on the table do not differ from any other table type.

2. Run a query from the TABLE_PARTITIONS table, as shown in Example 2-18, and you see that the new table features two partitions: one in the default store, and another in extended storage, as shown in Figure 2-19.

Example 2-18 SAP HANA Dynamic Tiering: Query SALES_ORDER table from the TABLE_PARTITIONS table

```
SELECT SCHEMA_NAME, TABLE_NAME, PART_ID, STORAGE_TYPE FROM TABLE_PARTITIONS
WHERE TABLE_NAME = 'SALES_ORDER' AND SCHEMA_NAME = 'DTUSER'
```

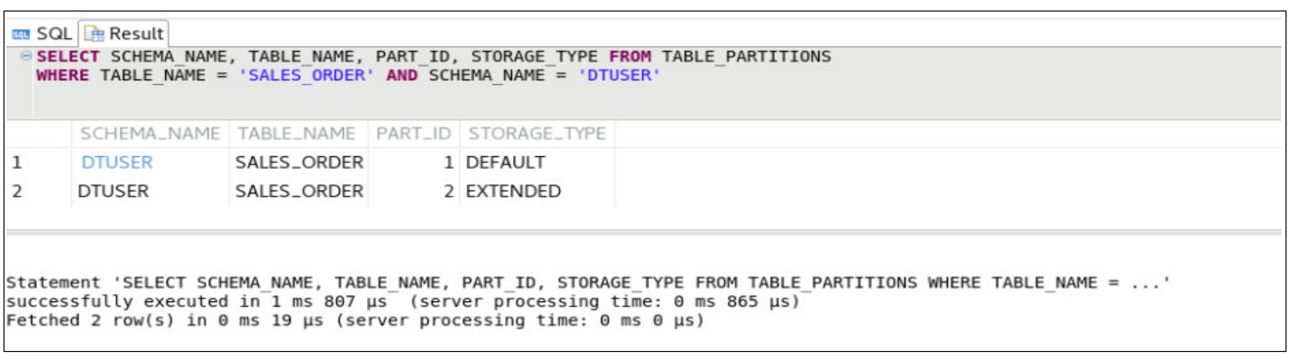


Figure 2-19 SAP HANA Dynamic Tiering: Results from TABLE_PARTITIONS table

3. Insert a row into the table so that it is stored in the default store by running the command that is shown in Example 2-19.

Example 2-19 SAP HANA Dynamic Tiering: Inserting a row in the default store partition of the SALES_ORDER table

```
INSERT INTO "DTUSER"."SALES_ORDER"
(S_SALESOKEY, S_CUSTOMER, S_VALUE, S_DATE)
VALUES
(1,1,120,'2011-12-11');
```

From table M_CS_TABLES (the table that shows data that is stored in the default store), which is shown in Example 2-20, you see that one record is in the default store as part of SALES_ORDER table, as shown in Figure 2-20.

Example 2-20 Checking the count of default store rows of the table SALES_ORDER

```
SELECT RECORD_COUNT FROM M_CS_TABLES WHERE TABLE_NAME = 'SALES_ORDER' AND
SCHEMA_NAME = 'DTUSER';
```

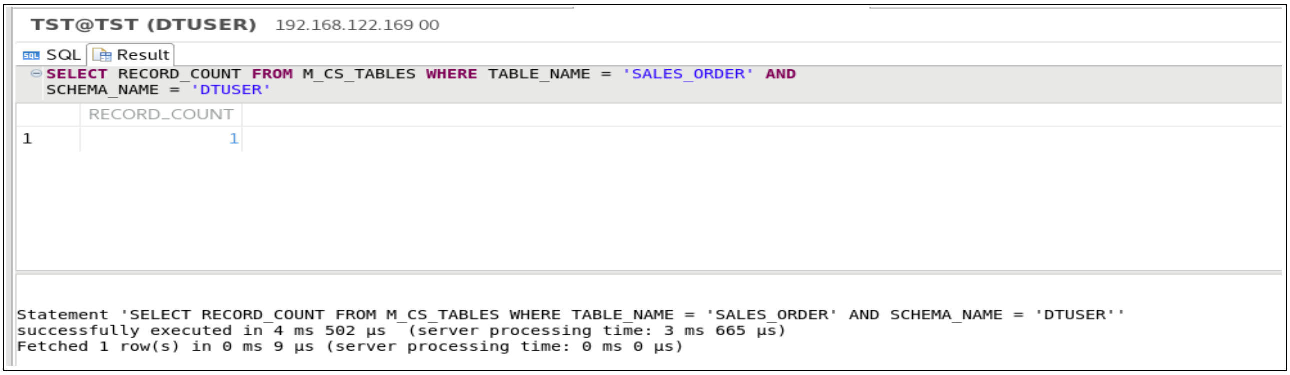


Figure 2-20 SAP HANA Dynamic Tiering: Results for the count of default store rows of the table SALES_ORDER

4. Insert one more row in the extended storage of the table SALES_ORDER, as shown in Example 2-21.

Example 2-21 Inserting a row in the extended storage partition of the SALES_ORDER table

```
INSERT INTO "DTUSER"."SALES_ORDER"
(S_SALESOKEY, S_CUSTOMER, S_VALUE, S_DATE)
VALUES
(1,1,120,'2009-12-11');
```

From the table M_ES_TABLES (the table that shows data that is stored in the extended storage) as shown in Example 2-22, you see that one record exists in the extended storage as part of SALES_ORDER table, as shown in Figure 2-21.

Example 2-22 Checking the count of extended storage rows of the table SALES_ORDER

```
SELECT * FROM M_ES_TABLES WHERE TABLE_NAME = 'SALES_ORDER' AND SCHEMA_NAME =
'DTUSER';
```

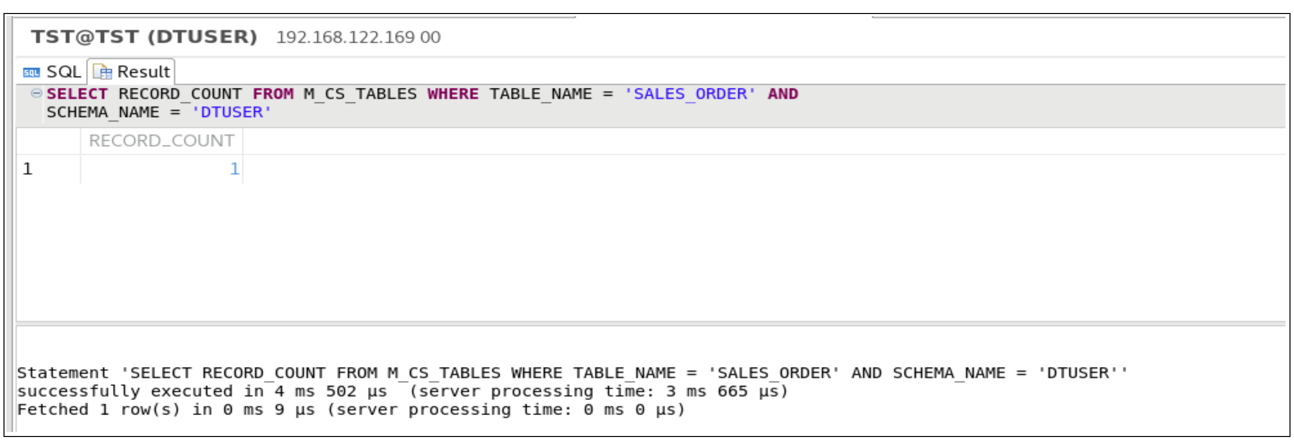


Figure 2-21 SAP HANA Dynamic Tiering: Results for the count of extended storage rows of the table SALES_ORDER

3

Fast-Start-Solutions for SAP HANA

As SAP HANA Online Analytical Processing (OLAP) databases (DBs) become larger, restarting them takes more time, which affects the availability of applications and the efficiency of your business.

This chapter describes different solutions that can help speed up starting large SAP HANA DBs to help minimize downtime.

This chapter includes the following topics:

3.1 Persistent memory and virtual persistent memory

In the quest to provide better performance for applications, new technologies across the computer industry are developed to help mitigate the inherent slowness of many of today's persistent disk-based storage solutions. IBM has a new PowerVM Persistent Memory architecture that is implemented at the hypervisor level. IBM also is developing several solutions to address this need.

3.1.1 Persistent Memory

A Persistent Memory is a type of memory with the following characteristics:

- Non-volatile: The ability to maintain contents after a power shutdown.
- Byte-addressable: The contents can be accessed by using CPU load and store instructions.
- Low latency: Refers to memory speeds, which are similar to that of Dynamic Random-Access Memory (DRAM).

At the same time that Persistent Memory dramatically increases systems performance, it enables a fundamental change in computing architecture.

Some SAP documentation refers to persistent memory as Non-Volatile Memory (NVM), while IBM Documentation often uses the term Storage Class Memory (SCM). The term Non-Volatile DIMM (NVDIMM) persistent memory also is used.

The Storage Networking Industry Association (SNIA) defined a programming model that describes an architecture of how operating systems can provide persistent memory services and how application software can use them. The PowerVM/Linux on Power Systems implementation of this programming model is shown in Figure 3-1 on page 37.

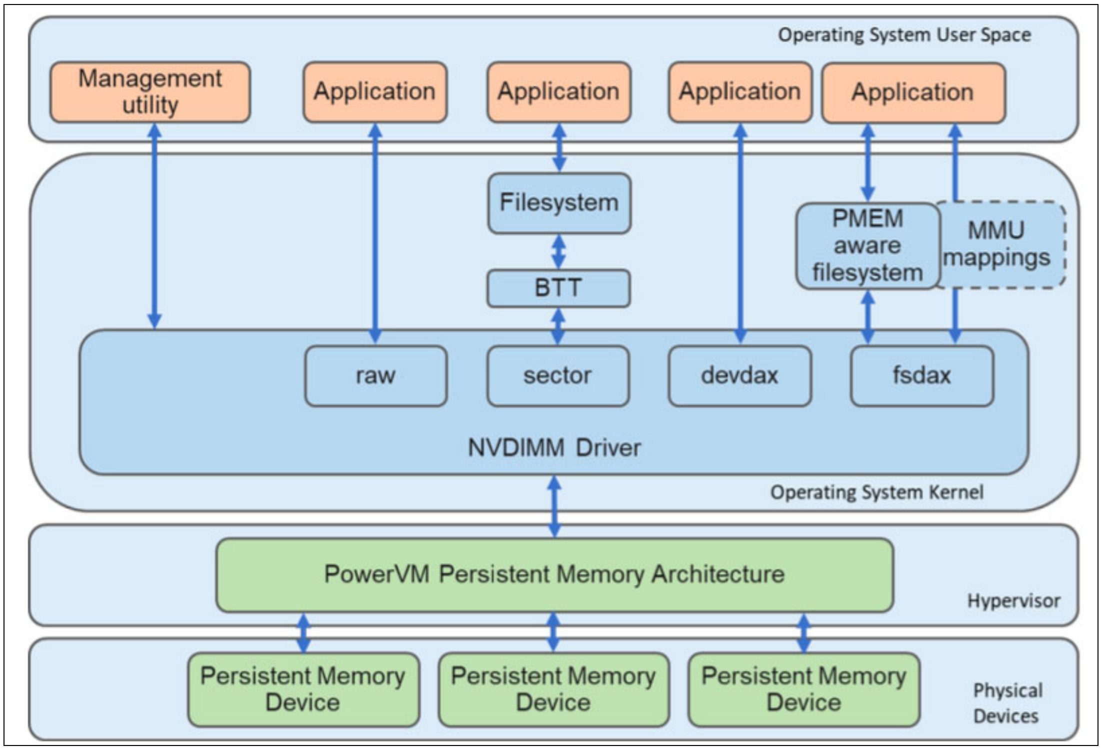


Figure 3-1 PowerVM / Linux persistent memory architecture

As shown at the bottom of Figure 3-1, the PowerVM hypervisor presents the persistent memory devices to the operating system in a technology agnostic manner. This process is referred to as the *PowerVM Persistent Memory Architecture*. This abstraction enables the adoption of new persistent memory technologies, attachments technologies, and device form factors with minimal effect on the operating system and virtualization management code.

Depending on the physical device capabilities, the PowerVM hypervisor is can virtualize persistent memory devices and segment them into smaller capacity volumes, which can be assigned to different logical partitions (LPARs).

After persistent memory is assigned to an LPAR, individual devices are presented by the Linux operating system as generic non-volatile DIMM devices, `/dev/nmem<#>`. The management tool `ndctl` is used to interface with the nvdimm driver to configure and provision these nvdimm devices into regions, namespaces, and persistent memory volume.

Region, which is grouping one or more NVDIMM devices, is commonly formed from devices from the same NUMA code.

Namespace is the partition of a whole or part of a region and is associated with a mode, which enables access methods to the persistent memory. The following modes are available:

- File system direct access (fsdax): Persistent memory is presented as a block device and supports XFS and EXT4file systems. This mode provides direct access (DAX) support, which bypasses the Linux page cache and performs reads and writes directly to the device. For direct access through load and store instructions, the device can be mapped into the address space of the application process with mmap(). The default mode of a namespace is fsdax.
- Device direct access (devdax): Persistent memory is presented as a character device. This mode also provides DAX support.

- Sector: Persistent memory is presented as a block device and supports any file system. This mode is useful for applications that are not persistent memory aware.
- Raw: This mode provides a memory disk with no DAX support.

For SAP HANA, only the fsdax mode is used. Figure 3-2 shows an example of the fsdax stack making available NVDIMM devices to application.

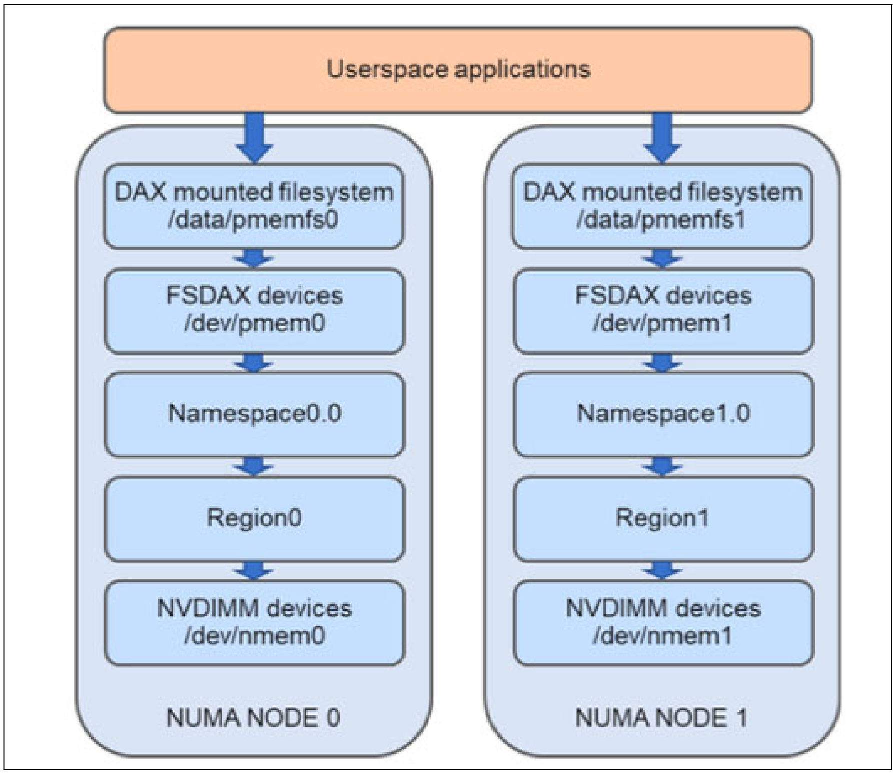


Figure 3-2 A fsdax mode stack

SAP HANA and Persistent Memory

SAP HANA uses persistent memory to reduce operational downtime. By retaining data in persistent memory after a shutdown, SAP HANA can avoid time-consuming data reloads from disk storage on start. For a large multi-terabyte SAP HANA database, this feature can reduce start time from well over an hour to only a few minutes, which is significant for systems that feature strict SLA requirements.

Specifically, SAP HANA supports placing column-store main data structures in persistent memory. The main data structures are highly compressed, read-only (after creation), and represent 95% of database data.

SAP HANA requires persistent memory to be configured in fsdax mode as shown in Figure 3-2 on page 38. Also, to take advantage of SAP HANA NUMA optimizations, the vPMEM volumes must be configured per NUMA node. File systems are created on the persistent memory fsdax devices and mounted by using the DAX option.

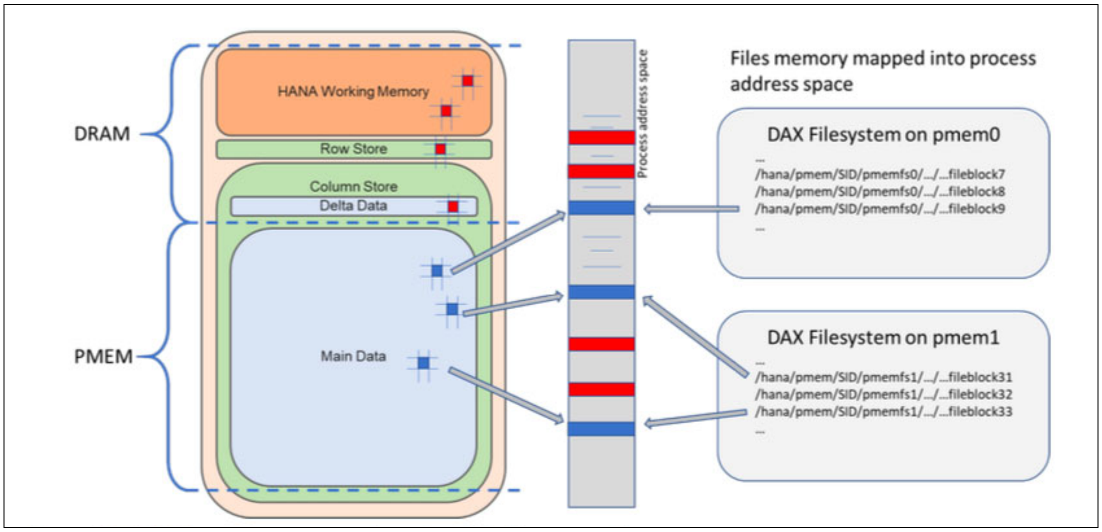


Figure 3-3 SAP HANA memory components and persistent memory data

SAP HANA main data, which is organized in column-wise data structures, can be written to files in the DAX file system. However, instead of using standard file I/O read and write calls, SAP HANA employs memory-mapped file I/O, as shown in Figure 3-3. By mapping the files directly into its address space, the application can use load and store CPU operations to manipulate the data.

3.1.2 Virtual persistent memory

A Virtual Persistent Memory (vPMEM) is a PowerVM feature that is offered on IBM POWER9 servers. It presents a portion of the installed standard system DRAM DIMMs as nvdimm devices to the operating system. The virtual qualifier denotes that this feature differs from true persistent memory considering DRAM is volatile memory. System DRAM loses its contents when the physical server is powered off.

This vPMEM technology is integrated into the IBM PowerVM hypervisor for POWER9 systems. It provides a high-speed persistent RAM disk storage solution for applications that persist across operating system and logical partition (LPAR) restarts.

However, powering down the physical system in a PowerVM virtualized environment is a relatively infrequent event. Maintenance is significantly more often performed at the level of LPAR logical partitions and shutdowns or restarts of the operating system do not involve powering down the physical server. As such, vPMEM nvdimm devices, also referred to as *vPMEM volumes*, maintain their contents over these operations.

Figure 3-4 summarizes the different levels of data persistence of vPMEM as compared to true persistent memory.

	vPMEM persistence	PMEM persistence
Application restart	✓	✓
LPAR restart	✓	✓
Physical server restart	✗	✓

Figure 3-4 vPMEM data persistence

The PowerVM Persistent Memory architecture allows for multiple types of memory to be defined and deployed for different use cases. Currently, the vPMEM solution creates persistent storage volumes from standard system DRAM, which provides high-speed persistent access to data for applications that are running in an LPAR.

For this solution, no special memory or storage devices are required; only unused available system DRAM is needed. It is intended that enhancements are to be supported that allow access to other types of memory to be used for different use cases.

vPMEM volumes are created as part of a specific LPAR definition and managed on the system Hardware Management Console (HMC). Each defined LPAR on a system can have a dedicated vPMEM volume.

Individual vPMEM volumes are not sharable between LPARs, and vPMEM volumes are not transferable to another LPAR, nor can be resized; instead, they are deleted and new vPMEM volumes can be created with the wanted size. They are sized on logical memory block (LMB) granularity where an LMB is the unit of memory that is used by the hypervisor to manage DRAM memory. By default, an LMB is 256 MB system-wide.

On creation, vPMEM volumes are specified to be striped across Non-Uniform Memory Access (NUMA) nodes or to be NUMA node contained. For NUMA aware applications, such as SAP HANA, vPMEM volumes are provisioned on a NUMA node basis so that their NUMA node associativity is clearly defined. That is, the PowerVM hypervisor allocates DRAM exclusively from one NUMA node to serve as single vPMEM volume, as shown in Figure 3-5 on page 41.

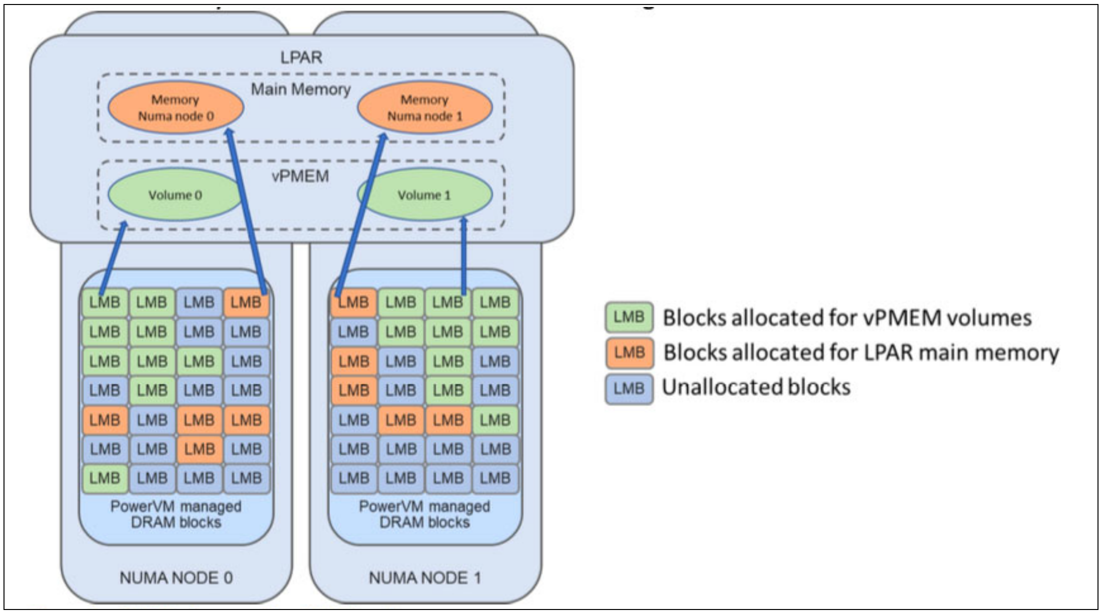


Figure 3-5 vPMEM placement by NUMA node

After the application uses this persistent system memory volume as a disk resource, any data that is stored in the vPMEM device persists if the LPAR is restarted.

Access to vPMEM volumes by the Linux operating system is provided by the standard non-volatile memory device (libnvdimm) subsystem in the Linux kernel corresponding **`ndctl`** utilities. The resulting vPMEM volumes are then mounted on the Linux file system as Direct Access (DAX) type volumes.

3.1.3 SAP HANA usage of vPMEM volumes

On POWER9 systems, SAP HANA v2.0 SPS04 revision 44 and later can use DAX-based persistent memory volumes to store column store table data for fast access. Persistent memory DAX file systems bypass the traditional file system page cache mechanisms, which increases the access speed to the data that is stored on those volumes compared to disk-based file systems.

When SAP HANA detects the presence of DAX vPMEM volumes, it starts the copy of main column store table data in these defined persistent volumes. Through its default settings, SAP HANA attempts to copy all compatible column store table data into the vPMEM volumes, which maintains a small amount of space in the LPAR DRAM for column store table metadata. This situation creates a persistent memory copy of the table data that SAP HANA then uses for query and transactional processing. SAP HANA can also be configured to copy in to vPMEM-only column store tables, or even specific partitions of individual column store tables.

To access the column store table data on the vPMEM volumes, SAP HANA creates memory-mapped pointers from the DRAM memory structures to the column store table data. Considering these vPMEM volumes are allocated from memory, accessing the table data is done at memory speeds with no degradation in performance compared to when the column store data is stored without vPMEM in DRAM.

Any changed or added data to tables that are loaded in to the vPMEM device is synchronized to disk-based persistent storage with normal SAP HANA save point activity. When SAP HANA is shut down, all unsaved data that is stored in the LPAR DRAM and the vPMEM volumes is synchronized to persistent disk.

When SAP HANA shuts down, the column store table data persists in the vPMEM volumes. The next time SAP HANA starts, it detects the presence of the column store table data in the vPMEM volumes and skips loading that data. SAP HANA re-creates memory structure pointers to the columnar table data that is stored on the vPMEM volumes, which results in a reduction in SAP HANA start times, as shown in Figure 3-6.

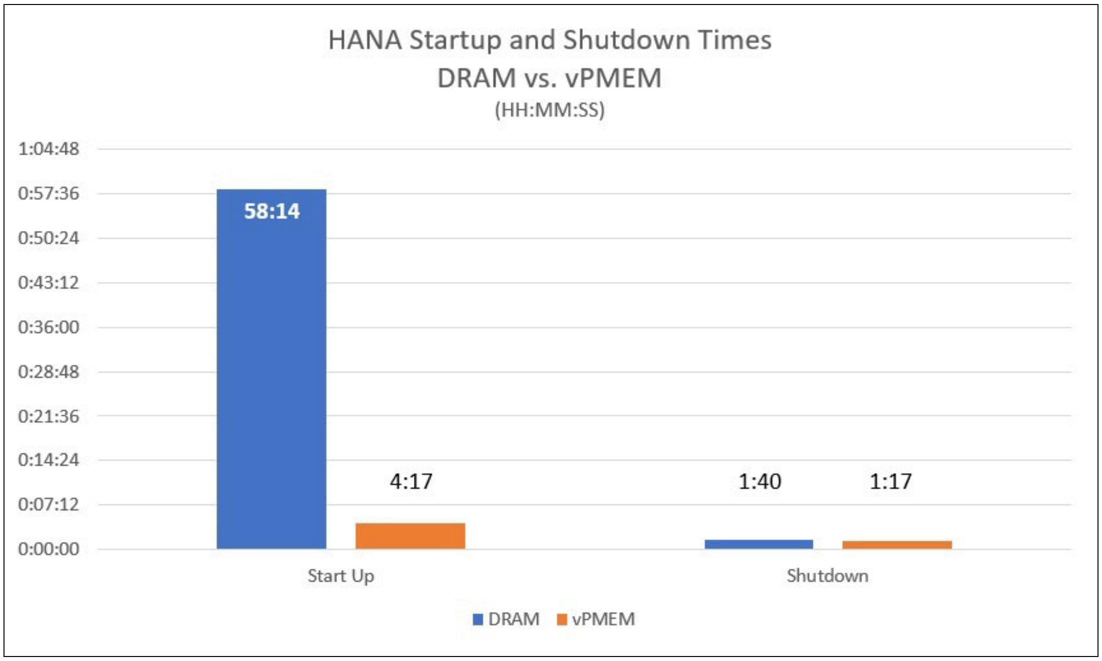


Figure 3-6 A large SAP HANA OLAP database start and shutdown times with and without vPMEM

This chart shows that substantial time savings for SAP HANA start for a large OLAP DB when all of the DBs column store data is allocated in vPMEM volumes. Time savings also are available for SAP HANA if shut down because not as much DRAM memory is required to be programmatically tagged as freed for ceding control to the operating system.

SAP HANA usage of vPMEM volumes

This section provides information about how to enable vPMEM usage for SAP HANA.

Prerequisites

The following minimum hardware and software levels are required to configure and implement SAP HANA with IBM PowerVM Virtual Persistent Memory:

- IBM POWER9 System with Firmware FW940.
- IBM HMC V9.1.940.
- SAP HANA 2 SPS04 revision 44.
- SLES 15 SP1:
  - Kernel version 4.12.14-197.21.1.
  - Ndctl version 64.1-3.3.1.

For more information about recommended patches, see SAP Note 2945828 - Virtual PMEM on IBM Power Systems.

To run the SAP Hardware and Cloud Measurement Tool (HCMT) with vPMEM, the minimum tool version is SAP HANA v2.0 SPS04 revision 46.

The next subsections of this chapter provide guidance about enabling vPMEM usage with SAP HANA.

3.1.4 Sizing the vPMEM volume for SAP HANA

Before configuring vPMEM volumes for use with SAP HANA in the Hardware Management Console (HMC), the suitable volume sizes must be determined. In general, vPMEM volumes are as large as the anticipated main data of SAP HANA system's column store plus capacity for growth and delta merge operations:

```
vPMEM Volume size = SAP HANA system's column store + capacity for growth +
capacity for delta merge operations
```

SAP Note 2786237 describes several tools to assist in the correct sizing of persistent memory:

- SAP HANA Quicksizer for greenfield deployments.
- Sizing report for SoH and S/4HANA (SAP Note 1872170).
- Sizing report for BWoH and BW/4HANA (SAP Note 2296290).
- SQL reports attached to the SAP Note 2786237 for an overview of memory usage in a current system.

> **Note:** The ratio restriction between DRAM and PMEM that is documented in SAP Note 2786237 does *not* apply to the POWER platform.

A rough estimation can be obtained by a simple query to the DB that provides the amount of memory in use by all columns store table (as shown in Example 3-1_ and adding 10 - 15% of memory to allow DB growth over time.

Example 3-1 Querying the amount of memory in use

```
select to_decimal(round(sum(memory_size_in_total/1024/1024),2,ROUND_UP),20,2)
COL_STORE_MB from m_cs_tables where schema_name = '<database schema name>'
```

If less LPAR DRAM memory is used by SAP HANA to store the column store table data that is on the vPMEM volume, the RAM memory allocation for the LPAR can be reduced by a similar amount to avoid the use of more system memory that is required for the LPAR. This memory definition adjustment is done in the LPAR profile on the HMC.

Checking the LPAR hardware assignment

It is important to have a vPMEM target LPAR assigned memory and CPU resources evenly across NUMA resource domains. SAP supports only LPAR and subsequent vPMEM allocations that are evenly distributed across the NUMA domains because PowerVM allocates memory resources to the LPAR that are in proportion to the CPU allocation across the NUMA domains.

If the LPAR is assigned CPU resources of 10 cores on one NUMA domain, say NUMA0, but only 5 on another domain (for example, NUMA1), the memory allocation follows this ratio, where the memory allocated on NUMA0 is twice the size as on NUMA1. This imbalance is also reflected in the size of the vPMEM volumes when they are created, which is a configuration that is not supported.

The resource allocation of CPU and memory for a system's LPARs can be queried by performing a *resource dump* from the HMC. The following methods are available:

- Within the HMC, which is described in How to Initiate a Resource Dump from the HMC - Enhanced GUI.
- By logging on to the LPAR's HMC command-line interface (CLI) with a privileged HMC user account, such as `hscroot`, and running the command that is shown in Example 3-2.

Example 3-2 Starting a resource dump from the HMC command-line interface

```
startdump -m <system name> -t resource -r 'hvlpconfigdata -affinity -domain'
```

Substitute <*system_name*> with the system name that is defined on the HMC that is hosting the LPAR.

Both methods create a resource dump file in the `/dump` directory that is timestamped. Depending on the size of the system, it can take a few minutes before the dump file is ready.

You can view the list of resource dump files on the HMC that is listed in chronological order by running the command that is shown in Example 3-3.

Example 3-3 Listing the resource dump files in the /dump directory

```
ls -ltr /dump/RSCDUMP*
```

The last file in the list can be viewed by running the **`less`** command, as shown in Example 3-4.

Example 3-4 Viewing the contents of the RSCDUMP file

```
Ò  less /dump/RSCDUMP.<system serial number>.<auto generated dump number>.<date
stamp>
```

The `less` command determines that the file can be in binary form because some of the data in the file is in this format. The details are in text form (see Example 3-5) for an IBM Power System E950 system with four sockets and 2 TB of RAM, with a single LPAR running that has been allocated 48 cores and 1 TB of the available memory.

Example 3-5 Main section of the RSCDUMP file list CPU and memory resources assigned to LPARs

```
|-----------|-----------------------|---------------|------|---------------|---------------|-------|
|  Domain   |    Procs       Units  |    Memory     |      |  Proc Units   |    Memory     | Ratio |
| SEC | PRI | Total | Free  | Free  | Total |  Free |  LP  |  Tgt  |  Aloc |  Tgt  |  Aloc |       |
|-----|-----|-------|-------|-------|-------|-------|------|-------|-------|-------|-------|-------|
|   0 |     |  1200 |     0 |     0 |  2048 |   311 |      |       |       |       |       |     0 |
|     |   0 |  1200 |     0 |     0 |  2048 |   311 |      |       |       |       |       |     0 |
|     |     |       |       |       |       |       |    1 |  1200 |  1200 |  1023 |  1023 |       |
|-----|-----|-------|-------|-------|-------|-------|------|-------|-------|-------|-------|-------|
|   1 |     |  1200 |     0 |     0 |  2048 |   269 |      |       |       |       |       |     0 |
|     |   1 |  1200 |     0 |     0 |  2048 |   269 |      |       |       |       |       |     0 |
|     |     |       |       |       |       |       |    1 |  1200 |  1200 |  1024 |  1024 |       |
|-----|-----|-------|-------|-------|-------|-------|------|-------|-------|-------|-------|-------|
|   2 |     |  1200 |     0 |     0 |  2048 |   312 |      |       |       |       |       |     0 |
|     |   2 |  1200 |     0 |     0 |  2048 |   312 |      |       |       |       |       |     0 |
|     |     |       |       |       |       |       |    1 |  1200 |  1200 |  1024 |  1024 |       |
|-----|-----|-------|-------|-------|-------|-------|------|-------|-------|-------|-------|-------|
|   3 |     |  1200 |     0 |     0 |  2048 |   311 |      |       |       |       |       |     0 |
|     |   3 |  1200 |     0 |     0 |  2048 |   311 |      |       |       |       |       |     0 |
|     |     |       |       |       |       |       |    1 |  1200 |  1200 |  1025 |  1025 |       |
|-----|-----|-------|-------|-------|-------|-------|------|-------|-------|-------|-------|-------|
```

In Example 3-5, the columns of data that are of interest in this context have the following meanings:

- Domain SEC

  The socket number in which the cores and memory are installed. In Example 3-5, the system features four sockets: 0 - 3.

- Domain PRI

  The NUMA domain number. Example 3-5 has four NUMA domains (0 - 3), and each NUMA domain aligns to a socket number. Some Power Systems servers have two NUMA domains per socket.

- Procs Total

  The number of processors in a 1/100th of a processor increment. Because PowerVM can allocate subprocessor partitions in 1/100th of a single core, this number is 100 times larger than the actual number of cores on the NUMA domain. Example 3-5 shows each socket features a total of 12 cores per socket.

- Procs Free/Units Free

  The total number of 1/100th of a core processor resource that is available.

- Memory Total

  The total amount of memory that is available on the NUMA domain. This number is four times larger than the actual memory in gigabytes that is available. Example 3-5 shows each socket has 512 GB of RAM installed, for a total system capacity of 2 TB.

- Memory Free

  The amount of memory that is not in use by assigned LPARs on the NUMA domain. Again, this value is four times larger than the actual memory in gigabyte of available memory. This measurement is an important detail in determining the amount of memory that can be used for a vPMEM volume because this value decreases after the creation of the vPMEM volume.

  Example 3-5 shows sockets 0, 2, and 3 all have approximately 75 GB of available memory, and socket 1 has about 65 GB of available memory. This system has a vPMEM volume of 700 GB that is assigned to the running LPAR.

- LP

  The LPAR number as defined in the HMC. Example 3-5 on page 45 shows only one LPAR running, LPAR number 1, which is assigned all CPU resources, and a subset of the available RAM resources.

- Proc Units Tgt

  The number of subprocessor processor units that is assigned to the LPAR from the NUMA domain. They are allocated by using the value from the Procs Total column. Example 3-5 on page 45 shows that the target allocation of processing units is 1200.

- Proc Units Aloc

  The number of subprocessor units that are allocated to the LPAR. Example 3-5 on page 45 shows that all 1200 units per socket are assigned and activated to the LPAR across all four NUMA domains or sockets.

- Memory Tgt

  The amount of memory that is assigned to the LPAR's DRAM configuration as defined in the LPAR profile. Again, this value is four times larger than the actual memory (gigabytes) that is assigned, and the hypervisor allocates this memory per NUMA domain in the same ratio as the processing unit assignment to the rest of the processing unit allocation across the other assigned NUMA domains.

  Example 3-5 on page 45 shows approximately 256 GB is targeted to be allocated to each NUMA domain, and in the same ratio as processing units. The memory is evenly distributed just as the processing units are evenly distributed.

- Memory Aloc

  The real allocation of memory to the LPAR per NUMA domain. Example 3-5 on page 45 shows all that memory that was requested is allocated to the LPAR. Summing up these values across the system reflects the LPAR DRAM memory allocation as seen by the operating system.

If the system has vPMEM volumes assigned, this memory allocation is not listed in this output. The memory values for the LPARs are the ones that are assigned to the LPAR's memory allocation in the profile. To determine the approximate amount of memory vPMEM volumes are taking on a specific socket, add up the memory allocations for the LPARs on that socket and subtract that value from the Memory Total. Taking this result and subtracting the value from Memory Free shows the amount of RAM that is used by the vPMEM volume, as shown in Example 3-6.

Example 3-6 General vPMEM memory allocation for a single socket

```
vPMEM memory allocation GB = ((Memory Total - sum(All LPAR MemoryAloc) - Memory
Free) / 4
```

Example 3-5 on page 45 for socket 0 used the values that are shown in Example 3-7.

Example 3-7 Calculating the vPMEM memory allocation for socket 0

```
socket0 vPMEM memory = (2048 - 1023 - 311) / 4 = 178. GB
```

Considering the same memory allocation is assigned across all four nodes, the total vPMEM device that is allocated to this LPAR is approximately 714 GB.

Managing and creating a vPMEM device

Configuration and management of virtual persistent memory (vPMEM) volumes is performed on an HMC that is running HMC V9.1.940 or later. The target POWER9 system needs firmware level FW940 or later.

vPMEM volumes are configured at the level of LPARs. Currently, creation, renaming, and deletion of vPMEM volumes is supported. To perform these operations, the LPAR must be in the state not activated.

To create a vPMEM memory device for an LPAR is done on the HMC by modifying the properties of the LPAR. Complete the following steps:

1. Click **System definition** in the HMC to get a list of available LPARs, and then, click the LPAR's name to get the general details of the partition. Then, click the **Persistent Memory** property to show the details for the vPMEM volumes, as shown in Figure 3-7.

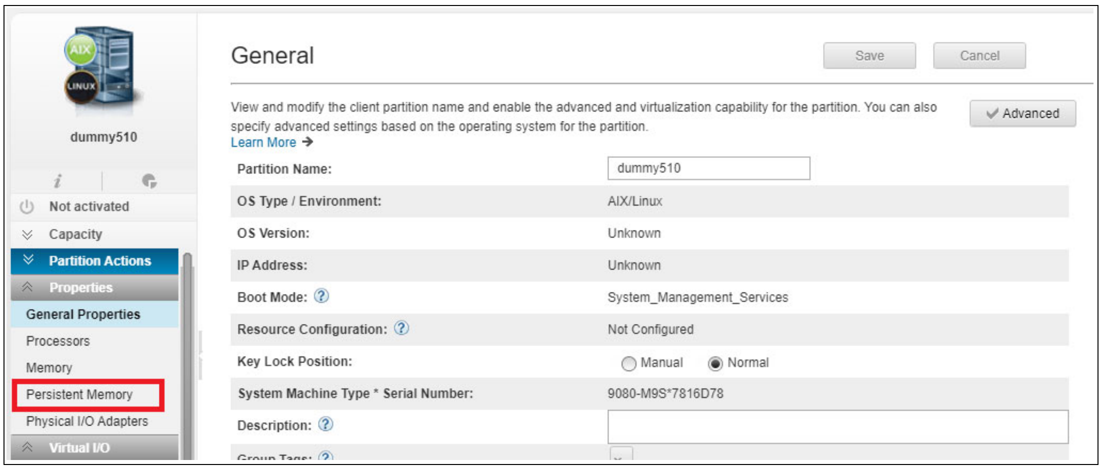


Figure 3-7 General properties: Persistent Memory options

By default, the list of vPMEM volumes is empty, as shown in Figure 3-8.

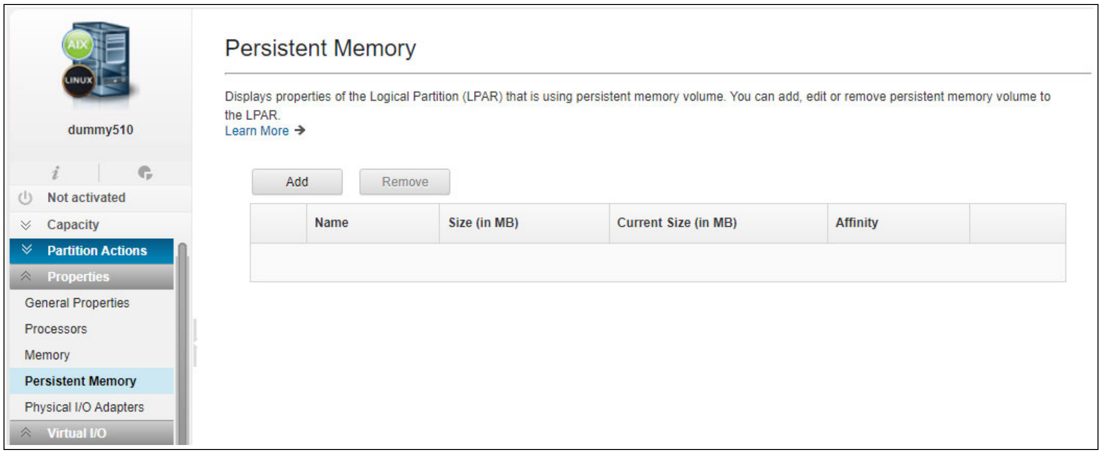


Figure 3-8 Persistent Memory window: Empty list of defined vPMEM volumes

2. Click **Add** to add a vPMEM device, as shown in Figure 3-9.

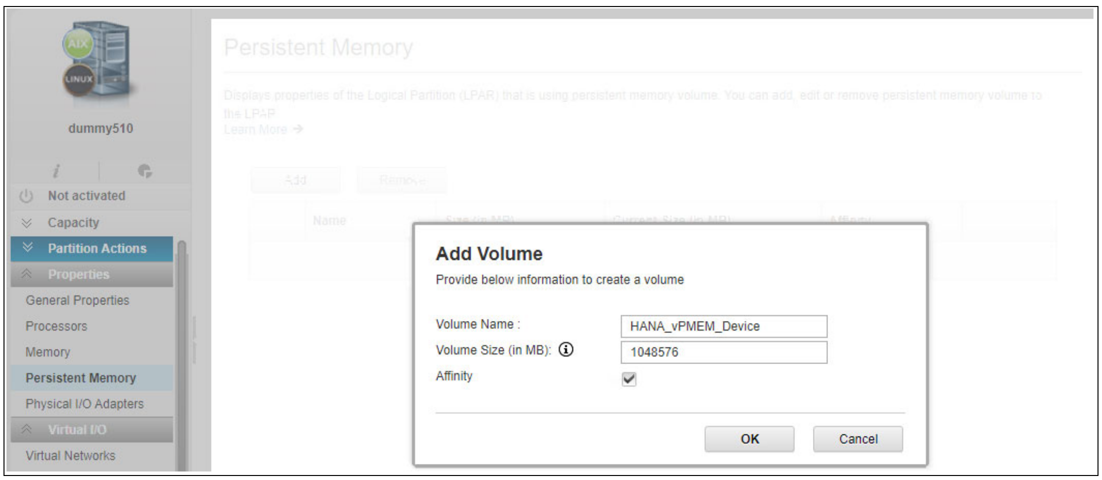


Figure 3-9 Adding a 1 TB vPMEM volume

3. Add a descriptive volume name, the total size in megabytes of the vPMEM device, and select the **Affinity** check box. Click **OK**, which creates a single vPMEM device for the LPAR.

After a vPMEM volume exists, you can rename or delete it on the Persistent Memory page.

In addition to the graphical configuration, the HMC command-line tools can be used. For example, use the **lshwres** command to list all vPMEM volumes of all LPARs on a managed system:

```
$ lshwres -r pmem -m ish359-HanaP-9009-42A-SN7800440 --level lpar
lpar_name=lsh30221,lpar_id=21,curr_num_volumes=0,curr_num_dram_volume
s=0,max_num_dram_volumes=4
lpar_name=lsh30222,lpar_id=20,curr_num_volumes=0,curr_num_dram_volume
s=0,max_num_dram_volumes=4
lpar_name=dummy2,lpar_id=5,curr_num_volumes=0,curr_num_dram_volumes=0
,max_num_dram_volumes=4
```

```
lpar_name=dummy1,lpar_id=4,curr_num_volumes=0,curr_num_dram_volumes=0
,max_num_dram_volumes=4
lpar_name=lsh30117-897d2131-
000001b2,lpar_id=3,curr_num_volumes=1,curr_num_dram_volumes=1,max_num
_dram_volumes=4
lpar_name=ish359v2,lpar_id=2,curr_num_volumes=0,curr_num_dram_volumes
=0,max_num_dram_volumes=4
lpar_name=ish359v1,lpar_id=1,curr_num_volumes=0,curr_num_dram_volumes
=0,max_num_dram_volumes=4
```

vPMEM affinity configuration options

When the vPMEM volume is configured for affinity (see Figure 3-10), the hypervisor divides the total allocated memory into equal parts that fit into the system memory of each of the LPAR's assigned NUMA memory nodes. From the operating system level, this division results in multiple `/dev/nmemX` devices that are defined, with one device for each NUMA node that is allocated to the LPAR.

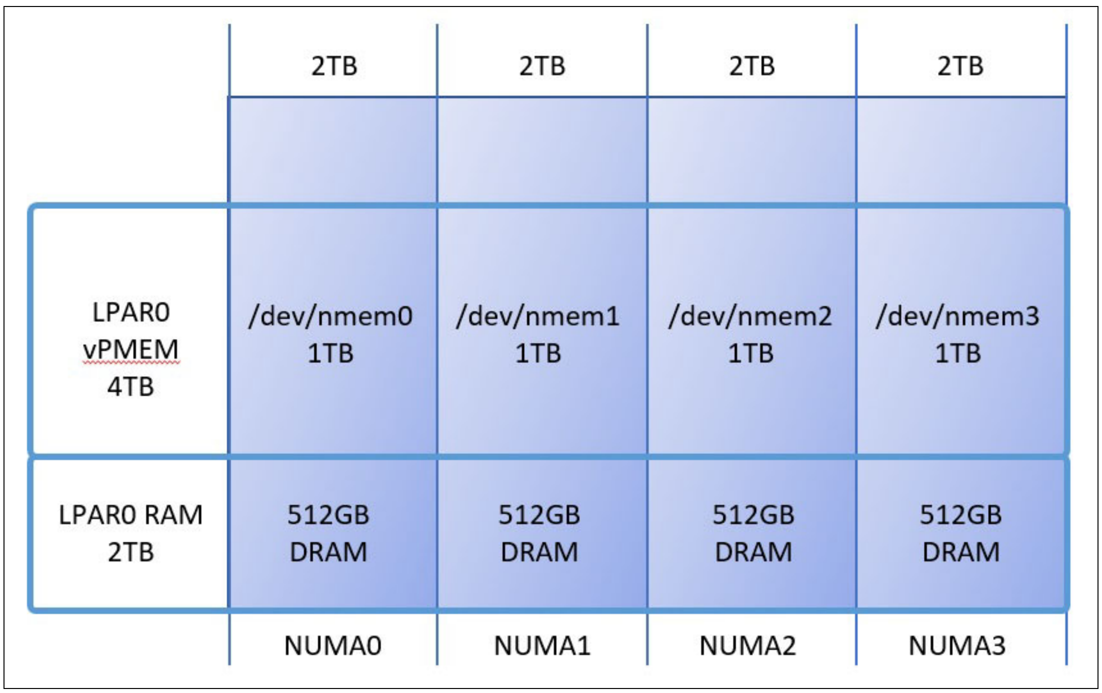


Figure 3-10 An 8 TB system with memory that is allocated across four NUMA domains

Figure 3-10 also shows an 8 TB system with memory that is allocated across four NUMA domains. Creating a 4 TB vPMEM device with NUMA affinity creates one vPMEM device per NUMA node, each 1 TB.

The benefit of dividing the vPMEM volume into segments and affinitizing them to NUMA boundaries enables applications to access data in physically aligned NUMA node memory ranges. Considering data is accessed sequentially, storing data NUMA optimized is best for throughput and access latency performance.

Affinitized vPMEM volumes are the only option that is supported by SAP HANA.

Affinity disabled

When the vPMEM device is activated without affinity, the hypervisor allocates a single vPMEM memory segment from the unused pool of system memory. It also divides this single memory region over all NUMA nodes to which the LPAR is assigned. When this single vPMEM device is activated at the operating system level, one `/dev/nmem` device is created, and all data that is copied to it also is divided over the NUMA nodes to which the LPARs are assigned. This information is less than desirable after all queries to the SAP HANA table data are retrieved across all NUMA nodes, which slightly increases latency of data access when compared to the affinitized option of the vPMEM volume creation.

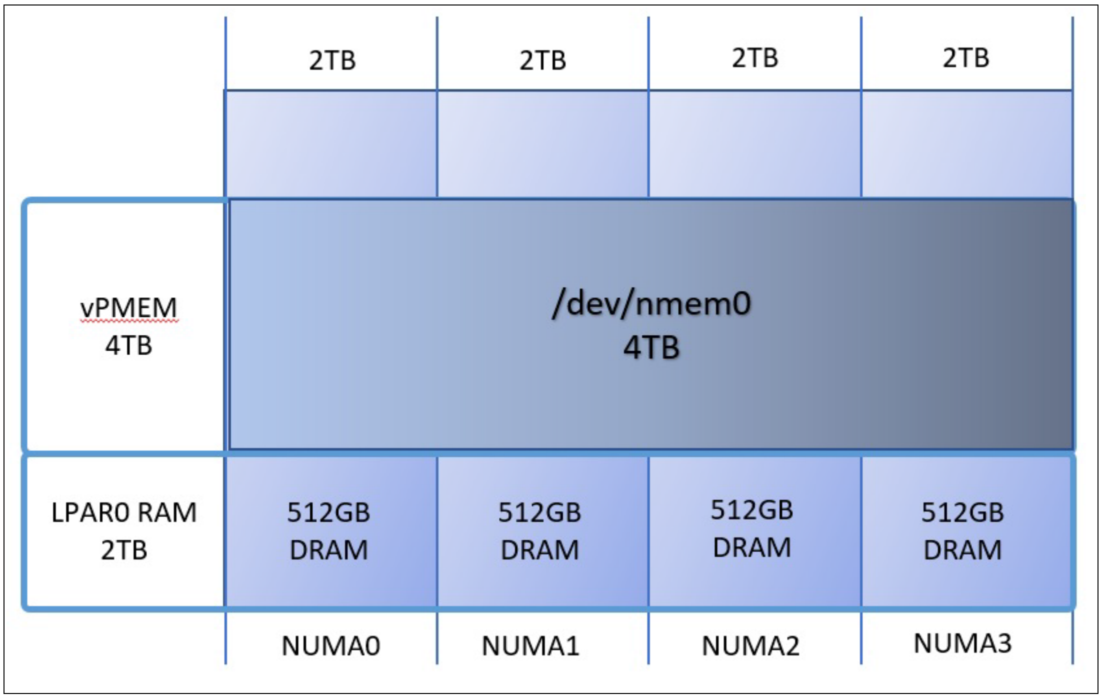


Figure 3-11 An 8 TB system with memory that is allocated across four NUMA nodes

Figure 3-11 also shows an 8 TB system with memory that is allocated across four NUMA nodes. This configuration creates a 4 TB non-affinitized vPMEM device that results in a single 4 TB device that is stripped across all NUMA nodes.

Currently, this vPMEM device option is *not* supported for SAP HANA.

Enabling vPMEM use in Linux

With Linux kernel 4.2, support for **`libnvdimm`** was introduced, which provides access to persistent memory volumes. When persistent memory volumes are activated by the operating system, raw memory devices are created in `/dev/nmem`, with one for each active persistent memory device. For Power Systems vPMEM volumes that are created with the Affinity option, one vPMEM volume is used for each NUMA node that is assigned to the LPAR.

The persistent memory volumes are then initialized, enabled, and activated with the standard operating system non-volatile DIMM control (**`ndctl`**) commands. Although these utilities are not provided by default in a base level Linux installation, they are included in the Linux distribution. Install them by running the application repository commands; for example, for Red Hat, install it by running the command that is shown in Example 3-8 on page 51.

Example 3-8 Red Hat command-line interface installation of the ndctl package

```
yum install ndctl
```

For SUSE Enterprise Linux Server, install it by running the command that is shown in Example 3-9.

Example 3-9 SUSE Enterprise Linux Server command-line interface installation of the ndctl package

```
zypper install ndctl
```

On Power Systems servers, each vPMEM volume is initialized and activated automatically. A corresponding number of `/dev/nmem` and `/dev/pmem` devices are available because NUMA nodes are assigned to the LPAR, as shown in Example 3-10.

Example 3-10 Listing of raw persistent memory devices

```
# ls -l /dev/*[np]mem*
crw------- 1 root root 241, 0 Jan 10 17:32 /dev/nmem0
crw------- 1 root root 241, 1 Jan 10 17:32 /dev/nmem1
crw------- 1 root root 241, 2 Jan 10 17:32 /dev/nmem2
crw------- 1 root root 241, 3 Jan 10 17:32 /dev/nmem3
brw-rw---- 1 root disk 259, 0 Jan 10 17:38 /dev/pmem0
brw-rw---- 1 root disk 259, 1 Jan 10 17:38 /dev/pmem1
brw-rw---- 1 root disk 259, 2 Jan 10 17:38 /dev/pmem2
brw-rw---- 1 root disk 259, 3 Jan 10 17:38 /dev/pmem3
```

If the `/dev/pmem` devices are not created automatically by the system during the initial operating system start, they must be created. Example 3-11 shows the set of **`ndctl`** commands to initialize the raw `/dev/nmem` device, where *X* is the device number (for example, `/dev/nmem0`).

Example 3-11 Creating /dev/pmem devices

```
# ndctl disable-region regionX# remove any previously defined regions
# ndctl zero-labels nmemX# clear any previously defined devices
# ndctl init-labels nmemX# initialize the new device
# ndctl enable-region regionX# enable the device region
# ndctl create-namespace -r regionX# create the region and /dev/pmem device
```

New device definitions are then created as `/dev/pmem`*`X`*. These new disk devices must be formatted, as shown in Example 3-12.

Example 3-12 Creating the XFS file system on the vPMEM device

```
# mkfs -t xfs  -b size=64k  -s size=512  -f /dev/pmemX
```

When mounting the vPMEM volumes, use the `/dev/disk/by-uuid` identifier for the volumes. These values are stable regarding operating system renaming of devices on restart of the operating system. Also, these volumes must be mounted by using the **`-o dax`** option, as shown in Example 3-13.

Example 3-13 Manual mounting of a vPMEM device to file system directory

```
# mount -o dax /dev/disk/by-uuid/34cb1120-1a61-47e5-9bcc-5b60e6d8e1d
/path/to/vPMEM/directory
```

To mount the volumes automatically on system restart, add an entry to the `/etc/fstab` file for each vPMEM volume by using the corresponding UUID name and adding the option **dax** in the options column. The use of the UUID name of the volume ensures correct remounting of the `/dev/pmemX` volume number after an operating system restart. Example 3-14 shows an entry for the `fstab` file for one vPMEM volume.

Example 3-14 Adding vPMEM devices into the /etc/fstab file

```
/dev/disk/by-uuid/34cb1120-1a61-47e5-9bcc-5b60e6d8e1d /hana/data/vPMEM/vPMEM0 xfs
defaults,dax 0 0
```

vPMEM volumes are not traditional block device volumes. Therefore, normal block device disk monitoring tools (for example, **iostat** and **nmon**), cannot monitor the I/O to the vPMEM devices. However, a normal director monitoring tool (for example, **du**) works because the files use the available storage space of the vPMEM volume.

Configuring SAP HANA to use vPMEM volumes

By default, SAP HANA usage of persistent memory volumes is specified at the host level. All HANA services that are managed by a single SAP HANA Global Allocation Limit (GAL) share a set of persistent Memory volumes.

Hence, the SAP HANA instance configuration must be updated to use the new vPMEM volume. Update the `global.ini` file to add the file system directory paths to the **basepath_persistent_memory_volumes** parameter in the `[persistence]` section, with each directory separated by a semi-colon, as shown in Example 3-15.

Example 3-15 Entry in the global.ini file defining the paths to the vPMEM volume directories

```
[persistence]
basepath_persistent_memory_volumes =
/path/to/first/directory;/path/to/second/directory;…
```

This parameter option is an offline change only, which requires the restart of SAP HANA to enable it.

On first start, SAP HANA by default copies all column store table data (or as much as possible) from persistent disk into the newly configured and defined vPMEM volumes. With partitioned column store tables, SAP HANA assigns in a round-robin fashion partitions to vPMEM volumes to distribute evenly the column store table partitions across the entire vPMEM memory NUMA assignment. When all column store data for a table is loaded into the vPMEM volumes, SAP HANA maintains a small amount of column store table metadata in normal LPAR DRAM memory.

Specifying tables, columns, or partitions to use vPMEM volumes

SAP HANA also can be configured to populate the vPMEM volumes with defined column store tables, columns, and partitions. First, change the `indexserver.ini` parameter to turn off the loading of all column store tables into the persistent memory, as shown in Example 3-16.

Example 3-16 Changing the default behavior of SAP HANA to not load all tables on SAP HANA start

```
[persistent memory]
table_default = OFF
```

The **`table_default`** parameter default value is `DEFAULT`, which is synonymous with the value of `ON`. This parameter, along with the **`global.ini`** parameter **`basepath_persistent_memory_volumes`**, makes the SAP HANA default behavior to load all table data in to the vPMEM devices.

The **`table_default`** parameter is dynamic. If a column stores table data in the vPMEM volumes, performing an SAP HANA UNLOAD of the unneeded columnar table data removes the table data from the persistent device. Then, a LOAD operation is needed to reload the table data into system DRAM, or a shutdown of SAP HANA can be done so that the old column store table data can be removed from the vPMEM volumes. On the next start, SAP HANA loads all column store table data into DRAM.

This setting can be overridden by the preference settings on the table, partition, or column level.

> **Note:** To specify different sets of vPMEM volumes for different SAP HANA tenants, use SAP Note 2175606 to first segment tenants to separate GALs. Then, define the persistent memory volumes in the `.ini` files at the database level.

Individual column store tables that are to use the vPMEM volumes can be moved by running the SQL command **`ALTER TABLE`**, as shown in Example 3-17.

Example 3-17 Altering a table move the column store table to PMEM

```
ALTER TABLE "<schema_name>.<table_name>" PERSISTENT MEMORY ON IMMEDIATE CASCADE;
```

This command immediately moves the table from LPAR DRAM to the vPMEM devices, and it persists for all future restarts of SAP HANA.

For columns and partitions, the only way to load data into vPMEM volumes is by running one of the **`CREATE COLUMN TABLE`** commands, as shown in Example 3-18.

Example 3-18 Specifying column store table columns or partitions in vPMEM

```
# create a table that the named column will use persistent memory
CREATE COLUM TABLE … <column> … PERSISTENT MEMORY ON

# create a table that the named partitions will use persistent memory
CREATE COLUMN TABLE … PARTITION .. PERSISTENT MEMORY ON
```

Column store table data can also be unloaded from the vPMEM devices and the data removed from the vPMEM volumes. Example 3-19 shows the commands that remove the column store table data from the vPMEM volume and unload the data from all memory.

Example 3-19 Altering a column store table to stop using vPMEM

```
ALTER TABLE <table_name> PERSISTENT MEMORY OFF IMMEDIATE
UNLOAD <table_name> DELETE PERSISTENT MEMORY
```

After this command runs, as shown in Example 3-19, the column store table data is no longer loaded into any memory areas (DRAM or vPMEM). Table data must be reloaded by future query processing or manually by running the SQL command that is shown in Example 3-20 on page 54.

Example 3-20 Reloading table data from disk-based persistent storage

```
LOAD "<table_name>" ALL
```

If the table persistent memory setting is changed to `OFF` or `ON`, it can be reset to `DEFAULT` by running the SQL **`ALTER TABLE`** command that is shown in Example 3-21.

Example 3-21 Changing the persistent memory setting

```
ALTER TABLE "<table_name>" PERSISTENT MEMORY DEFAULT
```

Verifying vPMEM usage

The following query can be used to verify that the vPMEM-based file systems are used by SAP HANA 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=30603
HOST,PORT,VOLUME_ID,NUMA_NODE_INDEX,PATH,FILESYSTEM_TYPE,IS_DIRECT_AC
CESS_SUPPORTED,TOTAL_SIZE,USED_SIZE
"lsh30117",30603,3,0,"/hana/shared/pmem/pmem0/JE6/mnt00001/hdb00003.0
0003","xfs","true",401517510656,15582494720
"lsh30117",30603,3,1,"/hana/shared/onen/pmem1/JE6/mnt00001/hdb00003.0
0003","xfs","true",402590728192,15930228736
```

The output shows that SAP HANA found and uses two persistent memory-based XFS file systems. One file system is backed by memory on NUMA node 0; the other is backed by memory on NUMA node 1.

Undersized vPMEM volumes

If the vPMEM volumes are too small to hold all of the column store table data, SAP HANA loads the remaining column store tabled data into LPAR DRAM. Queries still process normally, accessing data from DRAM or the vPMEM volumes. On an SAP HANA restart, SAP HANA loads the non-vPMEM stored column table data into LPAR DRAM from persistent disk.

Disabling the usage of vPMEM volumes

If it is necessary to stop using the vPMEM volumes, shut down SAP HANA and comment out the parameter **`basepath_persistent_memory_volumes`** and the vPMEM volumes in the `global.ini` file. On restart, SAP HANA loads all tables into the LPAR DRAM. The contents of the vPMEM volumes can then be cleaned out by reformatting them or deleting the contents from the directories. The vPMEM volumes then can be redeployed.

Benefits of the use of vPMEM volumes on Power Systems servers

The use of vPMEM volumes includes the following benefits:

- Faster restart of SAP HANA because column store table data is persistently loaded into vPMEM volumes.
- No special hardware is required to create and use vPMEM volumes for SAP HANA. The vPMEM volumes are created out of spare system memory.
- Access to data on vPMEM volumes is at RAM speeds, with no loss in performance compared to accessing table data in the traditional LPAR DRAM configuration.

- Multiple LPARs on a system can be assigned vPMEM volumes so that multiple applications on a system can use vPMEM and its benefits. Intel Optane is not supported in virtualized environments; therefore, only a single application per system is supported.
- SAP HANA attempts to align full columnar table data to a single vPMEM device, which aligns the table data to the memory of a single NUMA and speeds up access to that data.
- Partitioned table data is assigned in a round-robin fashion to NUMA nodes, which keeps all data within the partition range localized to a single NUMA node.
- After SAP HANA shuts down, deleting an vPMEM device does not affect any data in the database because SAP HANA synchronizes and saves new data to the persistent disk layer on shutdown.
- The maximum vPMEM segment per NUMA node is 4 TB.
- The vPMEM configured SAP HANA instances are supported and transparent to SAP HANA System Replication (HSR)/Distributed Statistics Records (DSR) configurations.

vPMEM usage considerations

Considering the following points about vPMEM:

- vPMEM device creation or deletion can be done only when the LPAR is inactive.
- vPMEM volumes are static in capacity. After they are created and assigned to an LPAR, the volume size cannot be changed.
- If a volume of a different size is required, the existing volume must be deleted, and a volume of a different size can be created.
- When you create a vPMEM volume, the new `/dev/nmem` devices must be reinitialized, and the resulting `/dev/pmem` devices must be formatted and mounted before SAP HANA starts. Then, SAP HANA repopulates the vPMEM volumes with columnar data during start.
- A defined vPMEM volume is assigned to a single LPAR and is not reassignable, which helps secure the data that is stored in the vPMEM volumes for the application that is running on the assigned LPAR.
- If the NUMA configuration for an LPAR that is assigned a vPMEM volume changes (for example, changes the number of cores), the previous vPMEM volume NUMA assignment does not match that of the new LPAR NUMA resource allocation. To maintain NUMA affinity, the vPMEM volume must be deleted and re-created to match the new NUMA alignment of the LPAR.
- vPMEM volumes at initial introduction are not Live Partition Mobility (LPM) capable. This feature is intended to be enabled as the technology matures.
- All `/dev/pmem` devices must be used directly. Partitioning these devices or creating LVM volumes is not supported because SAP HANA queries the device tree to understand the NUMA alignment of the devices. Any volumes that are created from the `pmem` device do not return valid NUMA alignment details, which cause SAP HANA to not start.
- SAP supports only one vPMEM device per NUMA node per SAP HANA instance.
- Multiple tenants of an SAP HANA instance can share vPMEM volumes. Each tenant must include a separate directory structure for the vPMEM volumes to be mounted.
- Only SAP HANA column store data is stored in vPMEM volumes. Row store tables and LOB column store tables are stored in DRAM memory.
- Log volume data is still stored on persistent disk for the protection of transactional data.
- Any changes to column store table data that is accessed from vPMEM volumes are protected by persistent disk through SAP HANA's save point synchronization mechanism.

PowerVM vPMEM enablement dependencies

To use vPMEM on Power Systems, dependencies exist on several components of the system, as shown in Table 3-1.

Table 3-1 Software version dependencies for using vPMEM on SAP HANA on Power Systems servers

Component	Version
Power Systems Firmware	Version 9.40 or later.
Hardware Management Console (HMC)	Version 9.1.940.0 or later.
Operating systems	SUSE Enterprise Linux Server 15 SP1 or later, or Red Hat 8.2 or later.
SAP HANA	Version 2.0 SPS4 rev44 or later.

vPMEM-related SAP Notes

For more information, see the following vPMEM-related SAP Notes:

- SAP Note 2700084
- SAP Note 2618154
- SAP Note 2786237
- SAP Note 2813454

Documentation for vPMEM and persistent memory usage

For more information about vPMEM and persistent memory usage with SAP HANA, see the following resources:

- IBM product documentation:

  Managing persistent memory volumes

- SAP product documentation:

  "Persistent Memory" in *SAP HANA Administration Guide for SAP HANA Platform*

3.2 RAM disk: SAP HANA Fast Restart Option

Temporary file systems (tmpfs) volumes are emulated file systems that are created from the available memory that is allocated to the operating system of an LPAR. All modern operating systems can create and use memory-based tmpfs volumes.

SAP HANA can use tmpfs volumes to store columnar table data in the same way it uses vPMEM volumes: The tmpfs memory volumes are mounted on the operating systems file system directory structure for storage of table data for fast access. *SAP* refers to the tmpfs volumes solution as the SAP HANA Fast Restart Option.

As with vPMEM persistent memory volumes, SAP HANA can use tmpfs volumes to store columnar table data in configured LPAR DRAM volumes. The mechanisms by which SAP HANA is configured to use tmpfs volumes are identical to vPMEM volume usage.

Like vPMEM volumes, access to the files that are stored in a tmpfs file system are performed at DRAM speeds. In this regard, accessing data from tmpfs and vPMEM volumes has the same performance. Also, no special hardware is needed to create tmpfs volumes because the DRAM memory that is allocated to the LPAR is used.

Unlike vPMEM, because the memory that is allocated to tmpfs volumes is allocated within the memory that is allocated to an LPAR, tmpfs volumes are not persistent through operating system or LPAR restarts. Data that is stored in tmpfs LPAR DRAM disks is volatile, and the contents are erased when the operating system is shut down or restarted. Upon restart of the operating system, the RAM disk volumes must be re-created, and SAP HANA columnar table data is reloaded from persistent disk volumes.

Also, unlike vPMEM volumes that are created with the Affinity option, creating a tmpfs volume is not automatically aligned to any specific NUMA node. NUMA node memory alignment details are gathered as a preparatory step and used in creating the tmpfs volumes at the operating system level.

One benefit of tmpfs file systems have over vPMEM volumes is that tmpfs volumes can be created to grow dynamically as the tmpf volumes fills, so the volumes can accommodate larger than expected data growth. But, this dynamic characteristic of tmpfs file systems includes the side effect of using more LPAR DRAM memory than expected, which funnels away memory from applications that need that memory to function. Hence, correctly sizing the tmpfs volumes is still important. Alternatively, the tmpfs volumes can be created to use a fixed amount of LPAR DRAM.

3.2.1 Fast-Start-Solution with TMPFS

This solution is based on temporary file systems that are defined on the operating system level. These temporary file systems are working like RAM based file systems. For each NUMA node one temporary file system must be created and mounted to a specific HANA directory. Figure 3-12 shows a scheme in which part of SAP HANA is to be stored in the temporary file systems.

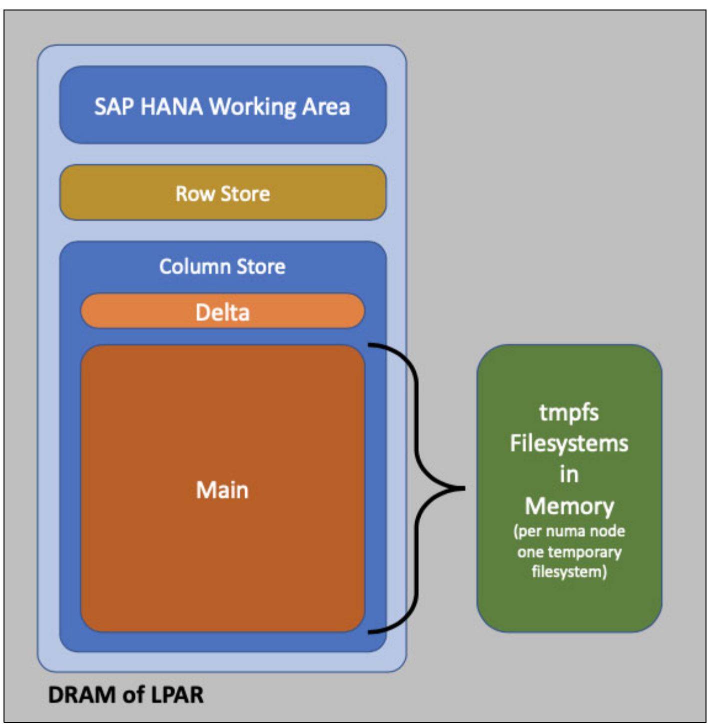


Figure 3-12 Scheme of tmpfs solution design

Configuring tmpfs volumes

Creating tmpfs file systems involves the use of LPAR DRAM memory to create volumes to store application data. Because memory is typically non-uniformly distributed across NUMA nodes, you must understand the memory architecture of the system before allocating memory to tmpfs file systems. This understanding ensures that the memory allocation of the tmpfs file systems can be allocated within NUMA nodes for fastest access.

A quick check at the operating system shows the available NUMA nodes and how much memory is allocated to each node, as shown in Example 3-22.

Example 3-22 Determining the amount of RAM that is allocated to each NUMA node of an LPAR

```
grep MemTotal /sys/devices/system/node/node*/meminfo
```

The command that is shown in Example 3-22 produces an output that shows the amount of memory that is available on each of the NUMA nodes that the operating system assigned.

Example 3-23 Example output from the grep MemTotal command

```
/sys/devices/system/node/node0/meminfo:Node 0 MemTotal:       510299264 kB
/sys/devices/system/node/node1/meminfo:Node 1 MemTotal:       510931712 kB
/sys/devices/system/node/node2/meminfo:Node 2 MemTotal:       511427328 kB
/sys/devices/system/node/node3/meminfo:Node 3 MemTotal:       511452992 kB
```

In this output (see Example 3-23), the system has four NUMA nodes, each installed with roughly 512 GB of system DRAM.

To create tmpfs devices in Example 3-23, allocate four different tmpfs devices, with one for each NUMA node. The **`mount`** command includes an option that can assign the memory for the tmpfs to a named NUMA node. Example 3-23 shows four directories to where the tmpfs files systems will be mounted.

Example 3-24 shows how to create the file systems by using the **`mount`** command options.

Example 3-24 Mounting the file systems

```
mount <tmpfs file system name>  -t tmpfs  -o mpol=prefer:X   /<directory to mount
file system>
```

In Example 3-24:

- `<tmpfs file system name>` is the operating device name. Use any descriptive name.
- `-t tmpfs` is the file system type, and in this case, of type tmpfs.
- `-o mpol=prefer:X` is the he NUMA node number to assign the memory for the tmpfs.
- `/<directory to mount file system>` is the location on the operating system file system path to mount the tmpfs file system. This directory is accessible and readable from the operating system level. Check that this directory exists as for any normal **`mount`** command.

In Example 3-23, the system has four NUMA nodes; therefore, four directories can be created and our different tmpfs file systems can be mounted (as shown in Example 3-25 on page 59) by substituting the `<SID>` with the SID of the SAP HANA DB.

Example 3-25 Script to create, mount, and make available tmpfs volumes for use by SAP HANA

```
for i in 0 1 2 3; do
   mkdir -p /hana/data/<SID>/tmpfs${i}
   mount tmpfs_<SID>_${i} -t tmpfs -o mpol=prefer:${i} /hana/data/<SID>/tmpfs${i}
done
chown -R <db admin user>:<db group> /hana/data/<SID>/tmpfs*
```

By using these options (see Example 3-25), the amount of memory that is allocated to each tmpfs is dynamically sized based on what is being stored by SAP HANA in the file systems. This option is preferred option because the file system grows as SAP HANA table data is migrated from RAM to the tmpfs files system.

If you must statically allocate an amount of memory to the tmpfs file system, use the **`-o size=<size in GB>`** option to allocate statically a fixed size of LPAR DRAM to the tmpfs file systems.

> **Note:** The tmpfs volumes are not formatted as an XFS file system like for vPMEM volumes, and the volumes are not mounted by using the **`-o dax`** option because of the differences in file system format for tmpfs volumes and the ability of SAP HANA to differentiate and support both types of file system formats to persistently store columnar table data.

Configuring SAP HANA to use the tmpfs volumes

Configuring SAP HANA to use tmpfs volumes is the same as with vPMEM memory:

1. Edit the `global.ini` file to specify the paths to the directories where the tmpfs volumes are mounted.
2. Verify that the `indexserver.ini` value for **`table_default`** is set to `DEFAULT` or `ON`.
3. Restart SAP HANA to populate the tmpfs volumes with column store table data.

> **Note:** More more set up information, see the SAP HANA Administration Guide for SAP HANA Platform - SAP HANA Fast Restart Option (which is available at this web page) and SAP NOTE 270084.

3.2.2 Fast-Start-Solution with vPMEM

The vPMEM solution is similar to the tmpfs solution for SAP HANA Fast-Start-Solution option. The main difference between the vPMEM and the tmpfs solution is the placement of the memory. The tmpfs Fast-Start-Solution uses the main memory of the LPAR; the vPMEM Fast-Start-Solution uses the persistent memory, which is defined for the LPAR.

In total, the summary of main memory and persistent memory features the same size as with the tmpfs or the Rapid-Cold-Start solution. Figure 3-13 on page 60 shows the architecture for vPMEM with SAP HANA on Power Systems.

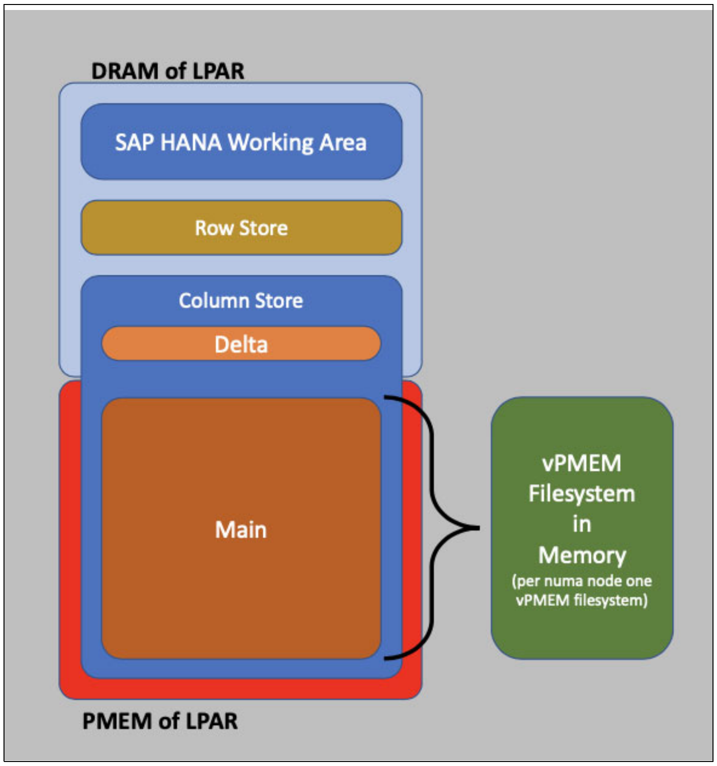


Figure 3-13 Scheme of vPMEM design

> **Note:** More details about the configuration and sizing for virtual persistent memory can be found at the following website:
>
> http://www.ibm.com/support/techdocs/atsmastr.nsf/WebIndex/WP102502

3.2.3 Comparing vPMEM and tmpfs

The use of tmpfs for holding column store table data has many of the same benefits as the use of vPMEM volumes. Table 3-2 lists some of the benefits of the use of persistent memory solutions on Power Systems servers.

Table 3-2 Benefits of using persistent memory solutions

Benefit	vPMEM	tmpfs
Faster restart of SAP HANA because column store table data is persistently loaded into mounted memory volumes.	Yes	Yes
Memory volumes persist across operating system and LPAR restarts.	Yes	No
No special hardware is required for SAP HANA to use mounted memory volumes for columnar table data.	Yes	Yes

Benefit	vPMEM	tmpfs
Access to data on mounted memory volumes is at RAM speeds with no loss in performance when compared to accessing table data in the traditional DRAM configuration.	Yes	Yes
Table data that is designated as non-preload on start is accessed at DRAM speeds, which increases query and application performance.	Yes	Yes
Multiple LPARs on a system can be assigned memory volumes.	Yes, and volume DRAM taken from spare system memory.	Yes, tmpfs volume created within each LPAR memory.
Automatic detection and alignment of memory volume to NUMA node.	Yes	Yes
SAP HANA attempts to align full columnar table data to a single memory volume.	Yes	Yes
Partitioned columnar table data is assigned in a round-robin fashion to memory volume.	Yes, with Affinity enabled.	Yes, when each file system aligned to NUMA nodes in the `mount` command.
After SAP HANA shutdown, deleting a vPMEM device does not affect any of the database data because SAP HANA synchronizes and saves new data to the persistent disk layer on shutdown.	Yes	Yes
Memory volume is Live Partition Migration (LPM) capable.	Not yet, but coming in a future release	Yes
Dependency on system firmware, HMC code, or operating system release.	Yes, as outlined in the previous section.	Can be used with any POWER9 supported versions.
SAP HANA Release Requirements.	>= 2.00.035	>=2.00.040

3.2.4 SAP-related documentation

For more information, see the following documentation:

- "Live Partition Migration" in *SAP HANA Administration Guide for SAP HANA Platform*
- SAP Note 270084

3.3 Persistent disk storage by using native Non-Volatile Memory Express devices or Fast-Start-Solution with Rapid-Cold-Start

This solution is not based on a technology that uses main memory compared to vPMEM and tmpfs. Instead, it uses NVMe devices to optimize the read performance of the attached I/O components.

Non-Volatile Memory Express (NVMe) adapters, with their high-speed flash memory, became a popular technology to store data that needs low latency and high throughput. The NVMe architecture and its access protocols became an industry standard, which makes it easy to deploy through a device driver addition to the operating system.

When you use NVMe in this context, you use the flash modules to store the DB persistently as though they are disk-based storage. SAP HANA reads data from these devices on start as they do from regular storage area network (SAN) -based storage. The benefit is that the read operations from NVMe are faster than SAP disk-based storage, which provides for faster SAP HANA start and for persistence of data across SAP HANA, operating system, and system restarts.

NVMe devices provide the following key benefits over SAN disk-based storage devices:

- Increased queue depth, which provides decreased I/O times.
- Lower latencies for read and write operations.
- Higher I/O throughput than traditional disk fabric systems (for example, Fibre Channel) because of the location of the adapters on a PCIe adapter slot.

Figure 3-14 shows the design of the Rapid-Cold-Start Solution.

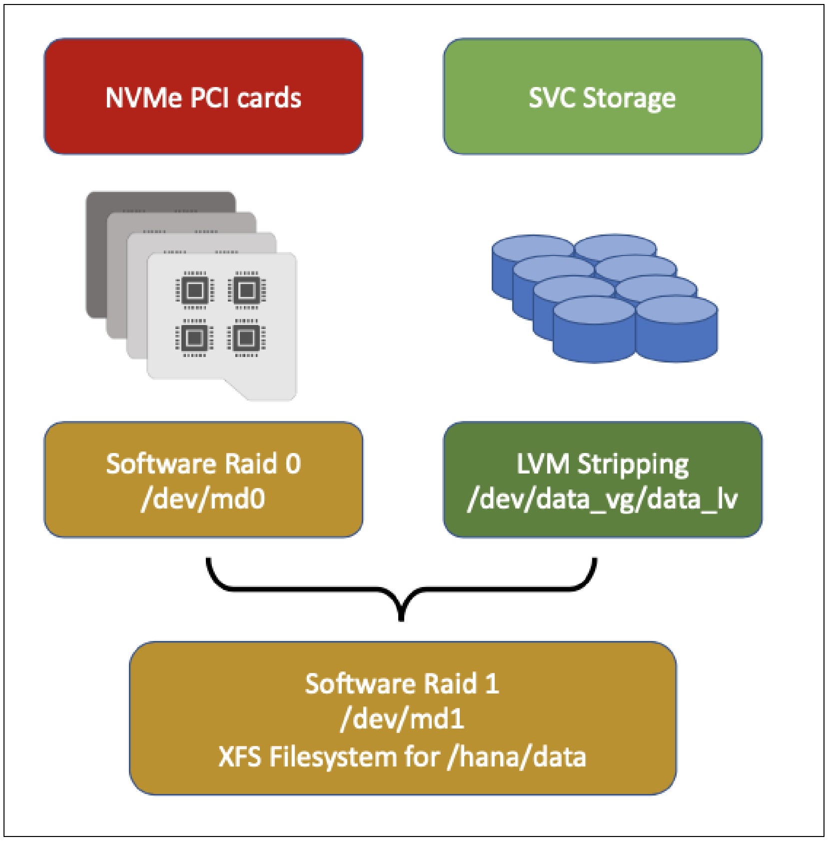


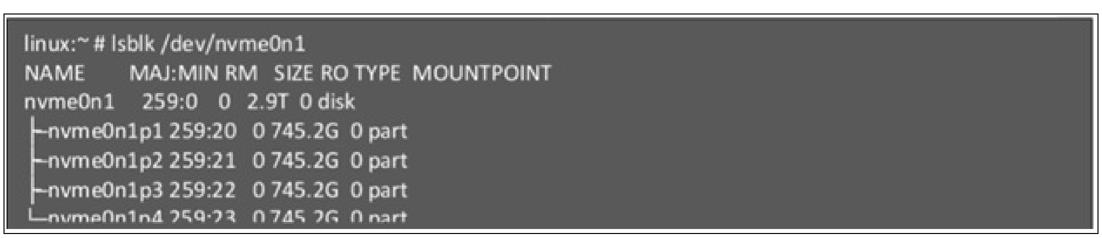
Figure 3-14 Scheme of the Rapid-Cold-Start solution

3.3.1 NVMe device details

NVMe devices are presented to the operating system through the traditional block device interface. For Linux, these devices are listed in the `/dev` directory along other disk-based block devices. Therefore, NVMe devices can be inserted into an LPAR as regular storage devices.

NVME adapters are made up of multiple individual flash memory modules. This architecture allows the operating system to access multiple storage modules per NVMe adapter independently.

Figure 3-15 shows a sample output that lists the devices on an individual NVMe 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
└─nvme0n1p4 259:23   0 745.2G  0 part
```

Figure 3-15 NVMe adapter module listing

Figure 3-15 also shows that four modules with 745 GB per module are used, and 3 TB total storage for the adapter.

3.3.2 Configuration and usage considerations

The flash NVMe adapters that are assigned to the LPAR can be used in several different ways:

- Directly used as an independent storage device.
- RAID arrays can be created to span multiple modules on an individual NVMe adapter.
- RAID arrays can span modules across other NVMe adapters.

Figure 3-16, Figure 3-17, and Figure 3-18 show a few examples.

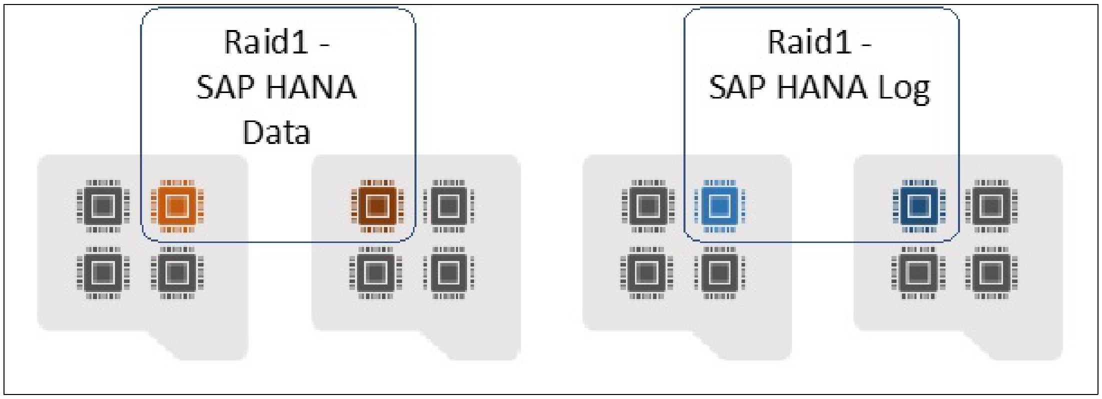


Figure 3-16 RAID 1 storage volumes for SAP HANA data and logs by using NVMe adapters

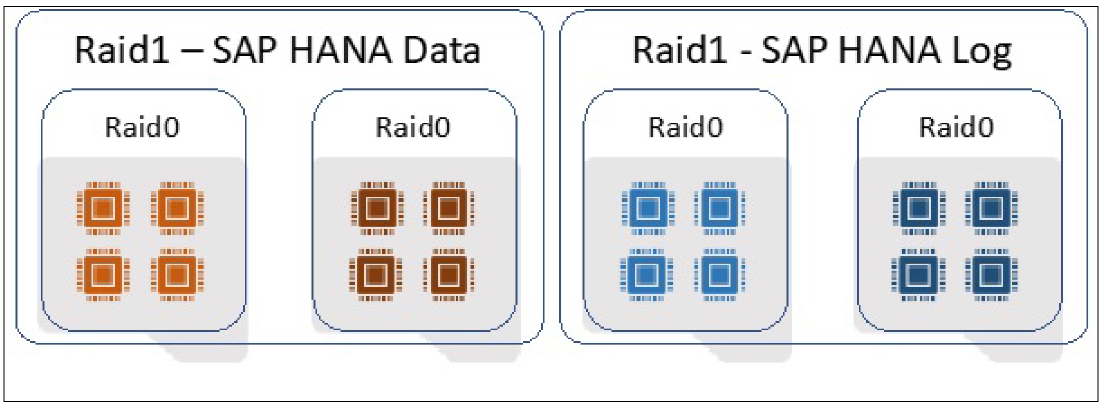


Figure 3-17 RAID 1 arrays for SAP HANA data and logs that are created

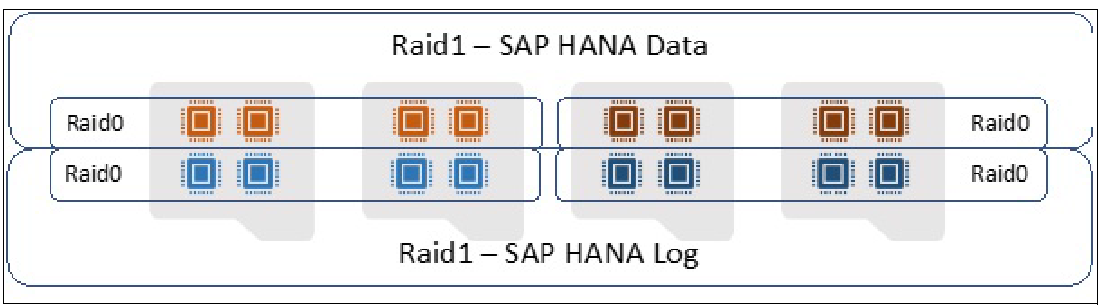


Figure 3-18 RAID 0 and RAID 1 arrays for SAP HANA data and logs

Because flash modules are subject to long-term degradation in high-write and data-change environments, be careful about which data is assigned to flash storage volumes. Preference must be given to data profiles that do not include excessive data modification rates.

For environments that must use flash modules with high-write and change activity characteristics, mirroring a NVMe volume with a traditional disk-based storage volume can help preserve data integrity if a NVMe flash wear fails.

Note: Monitoring and alerting muse by set up in case a card fails to instantly take corrective actions.

3.3.3 NVMe performance characteristics

Tests that are conducted by using SAP HANA and high I/O workloads show that the use of NVMe adapters for SAP HANA volumes can decrease latency by up to 9x over traditional storage technologies (for example, IBM SAN Volume Controller that uses 20 GB cache and IBM XIV® storage).

For read performance, testing shows that NVMe adapter volumes are up to 2 - 4x faster than disk-based storage solutions (depending on I/O block sizes) as shown in Table 3-3.

Table 3-3 NVMe performance characteristics

Block size	64 KB	128 KB	256 KB	1 MB	16 MB	64 MB
Performance increase by using NVMe	2.2x	2.7x	4.0x	4.5x	4.0x	4.1x

Write performance also increases by up to 2x, as shown in Table 3-4.

Table 3-4 VVMe write performance characteristics

Block size	4 KB	16 KB	64 KB	128 KB	256 KB	1 MB	16 MB	64 MB
Performance increase initial write while using NVMe	N/A	1.1x	1.4x	1.7x	1.9x	1.8x	1.7x	1.7x
Performance increase overwrite while using NVMe	1.5x	1.8x	2.0x	1.8x	1.7x	1.9x	1.6x	1.6x

3.3.4 Striping effects for internal NVMe cards

Creating a RAID 0 volume over multiple NVMe adapters increases the throughput on read operations on block sizes greater than or equal to 256 KB nearly by a factor of 2.

Creating a RAID 0 volume over multiple NVMe devices increases the throughput on write operations on block sizes greater than or equal to 64 KB nearly by a factor of 1.7. On block sizes greater than 256 KB, the factor is nearly 2.

Creating a RAID 0 volume on the multiple memory modules of one NVMe device has no positive effect on performance over storage on a single memory module.

3.3.5 Summary

NVMe adapters with their fast flash modules can provide an increase in I/O performance for SAP HANA compared to storage on traditional SAN storage options, which can be solid-state drives (SSDs) or hard disk drives (HDDs). The increase in speed and lower latencies is provided by much faster flash technology in the NVMe adapter versus SSD flash storage and by accessing the NVMe storage through PCIe 3.0 protocol versus over Fibre Channel protocols.

3.4 NVMe Rapid-Cold-Start Mirror

Another option for NVMe adapter flash storage is to mirror the volumes to external SAN disk-based persistent storage for high availability (HA), and faster start of large SAP HANA DBs.

A large DB takes some time to load from SAN disk-based storage through connectivity solutions, such as Fibre Channel that are connected to HDD-based or SSD-based volumes. NVMe storage that is hosted at the host level provides faster access to the SAP HANA data for loading and writing.

However, flash modules are subject to wear because of data writes or changes. To protect valuable data, a hybrid of the traditional SAN disk-based storage and NVMe storage can be used to provide faster SAP HANA start with protection of the data on disk-based storage.

3.4.1 Hybrid mirrored volume

The solution that is shown in Figure 3-19 creates RAID 0 arrays from the NVMe FLASH modules, and a RAID 1 mirror to the storage that is created on the SAN disk-based storage solution.

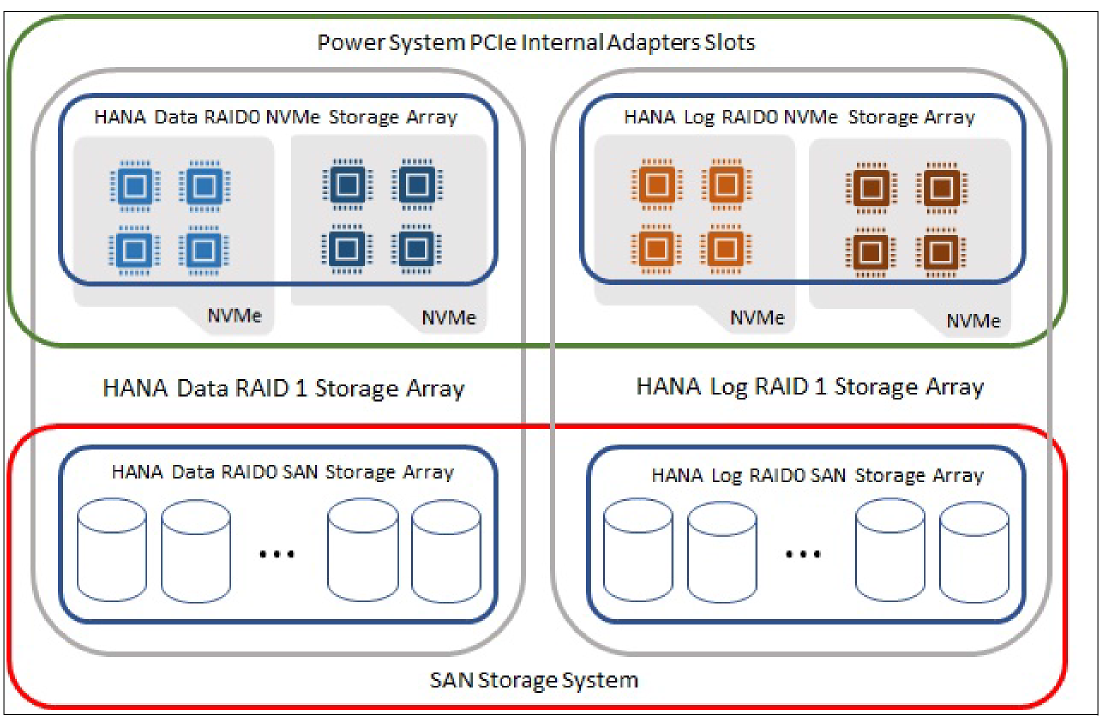


Figure 3-19 NVMe and SAN disk-based storage mirror

In Figure 3-19 on page 66, the NVMe adapter flash modules (inside the green box) are all added into a RAID 0 array. On the SAN side, RAID 0 arrays of the same size as the NVMe RAID 0 array are created (inside the red box). Each of these four RAID 0 volumes are represented with blue boxes on their respective storage platforms.

Then, a mirrored RAID 1 volume is created by assigning one SAN RAID volume to one NVMe RAID 0 volume, which is represented by the gray boxes in Figure 3-19 on page 66.

When creating the RAID 1 mirror between the NVMe volumes and the SAN storage volumes, you can set the preference for the operating system to write to the NVMe volumes by passing the option **`--write-mostly`** of the **`mdadmin`** array utility. In this case, the preference is to assign it to the RAID 0 device name for the external SAN volume. Example 3-26 shows the Linux man page excerpt for **`mdadm`**.

Example 3-26 The mdadm --write-mostly option to prefer one RAID device for reading

```
-W, --write-mostly
   subsequent devices listed in a --build, --create, or --add command will be
   flagged as 'write-mostly'.  This is valid for RAID 1 only and means that the
   'md' driver will avoid reading from these devices if at all possible.  This
   can be useful if mirroring over a slow link.
```

In this case, for the RAID 1 device, specify the SAN storage device as **`--write-mostly`** in the **`mdadm --create`** command, as shown in Example 3-27.

Example 3-27 Specifying the SAN storage device

```
mdadm --create --verbose --/dev/<new RAID device name> --level=1 --raid-devices=2
/dev/<NVMe device name --write-mostly /dev/<SAN storage device name>
```

3.4.2 Summary

Mirroring an internally installed NVMe adapter to an external SAN volume of the same size can provide the benefits of a rapid start of SAP HANA by reading data from the NVMe adapter with the RAID 1 mirroring of the data to an external SAN disk for ultimate data protection.

3.5 Comparing vPMEM to Intel Optane

The Optane persistent memory solutions that are available on the Intel platform constrict technology to the persistent memory solutions vPMEM and tmpfs for fast restart. The Optane memory Data Center Persistent Memory Module (DCPMM) solution is based on a new memory DIMM architecture that allows for persistent storage of data that is stored to the DCPMMs, which allows that data to survive operating system and system restarts.

vPMEM and tmpfs for fast restart also provide resiliency of the data from an SAP HANA restart. In addition, for vPMEM for an LPAR restart, data that is stored in these solutions must be restored in a full system restart.

Optane DCPMM memory is implement by installing new Optane DCPMM memory cards into existing system DRAM DIMM slots. Real DRAM capacity must be sacrificed to use the Optane memory solution.

In contrast, the vPMEM option for Power Systems serves uses standard system DRAM that is installed in the system. Implementation of vPMEM is as simple as defining the vPMEM memory segment, starting the LPAR, and configuring SAP HANA to use the persistent memory segments. No hardware downtime is necessary unless more DRAM is required to support the vPMEM solution.

Optane memory capacities are provided in 128 GB, 256 GB, and 512 GB sizes per PCDMM module, which is much higher capacities than what exists for DIMM memory modules that have a maximum capacity of 64 GB per DIMM.

Rules for the use of PCDMM modules are complicated and vary depending on the memory mode that is used. However, with a maximum of 12 DIMM modules per socket that uses six DIMM modules of 64 GB and the maximum of six 512 GB DCPMM memory modules, a socket maximum memory configuration is 3.4 TB. This memory configuration is compared to a maximum memory configuration of 1.5 TB when only DIMM memory modules are used. POWER9™ systems support up to 4 TB of DIMM system memory per socket by using 128 GB DIMMs. Future DIMM sizes will increase this memory footprint.

From a memory latency perspective, Optane DCPMM memory modules have a higher read and write latency compared to standard DIMM memory technologies because of the technology that is implemented to provide data persistence in the DCPMM module. Higher latencies can affect application performance and must be evaluated when implementing the Optane solution.

IBM POWER9 vPMEM and tmpfs use DIMM-backed RAM and perform read and write operations at full DIMM throughput capabilities.

Optane has three memory modes that the DCPMM modules can use:

- Memory Mode

  In this mode, the DCPMM memory modules are installed along standard DIMM modules and are used as a regular memory device. One advantage of the use of the DCPMM modules in this mode is that greater overall memory capacities can be achieved over standard DIMM modules that are available for x86 systems.

  However, enabling the PCDMM modules in this mode puts the regular DIMMs in the system into a caching function, which makes their capacity invisible to the host operating system. Therefore, only the capacity of the DCPMM memory can be used by the host operating system, and the regular DIMM memory capacity is unavailable for operating system and application use.

- App Direct Mode

  In this mode, the DCPMM memory modules are used as persistent storage for operating systems and applications that can use this technology. The DCPMM memory modules are recognized by the operating system as storage devices and are used to store copies of persistent disk data, which makes access to that data faster after an operating system or system restart. The standard DIMMs are used normally as available RAM to the operating system.

- Mixed Mode

  This mode is a mixture of the Memory and App Direct modes. When DCPMM modules are used in this mode, a portion of the capacity of the module is used as memory for the host operating system, and the remaining capacity is used for persistent storage. However, as in Memory Mode, any DIMM memory is unavailable for use by the host operating system and is instead converted into a memory cache subsystem.

Currently, Optane persistent memory is not supported for use by SAP HANA in virtualized environments.

3.6 Scenario comparison between the different Fast-Start-Solutions

During the lifetime of an SAP HANA system, many different triggers can be used for starting and stopping the SAP HANA and the underlying operating system. Figure 3-20 lists these different triggers and the effect for the different Fast-Start-Solutions.

Scenario	tmpfs	vPMEM	Rapid-Cold-Start	Internal NVMe Cards	FS9100 NVMe Storage
SAP HANA start (initial)	no	no	yes	yes	yes
SAP HANA start after stop	yes	yes	yes	yes	yes
SAP HANA restart	yes	yes	yes	yes	yes
SAP HANA service restart	yes	yes	yes	yes	yes
Operating system reboot	no	yes	yes	yes	yes
Operating System update/service patch	no	yes	yes	yes	yes
System restart	no	no	yes	yes	yes

Figure 3-20 Different scenarios/triggers for start and stop action

3.6.1 Application memory comparison

The application performance differences between the Fast-Start-Solutions goes to zero. Figure 3-21 shows the memory configuration and memory usage by SAP HANA on the tested scenarios.

	Standard	tmpfs	vPMEM	NVMe
Total Memory	4091 GB	4091 GB	4006 GB	4091 GB
No Persistent memory	4091 GB	4091 GB	511 GB	4091 GB
Persistent memory	0 GB	0 GB	3495 GB	0 GB
Memory by SAP HANA	2921 GB	2923 GB	2924 GB	2924 GB

Figure 3-21 Memory values

3.6.2 Performance differences between the Fast-Start-Solution variants

Figure 3-22 lists the measurement data of the start and stop tests for the benefit validation of the different Fast-Start-Solution methods for SAP HANA. This table considered only the start and stop behavior of SAP HANA. The start and stop times for LPAR operations or system operations are not included in this table.

	Standard	tmpfs	vPMEM	Rapid-Cold
Start times after System Power Off Benefit against standard	00:35:45 -	00:29:00 x 1,2	00:31:02 x 1,2	00:05:19 x 6,5
Start times after LPAR down Benefit against standard	00:32:13 -	00:28:51 x 1,1	00:02:27 x 13	00:05:00 nearly x 6,5
Start times after SAP HANA Stop (table load = 5 parallel) Benefit against standard	00:31:44 -	00:02:22 x 13	00:02:24 x 13	00:04:59 nearly x 6,5
Start times after SAP HANA stop (table load = 12 parallel) Benefit against standard	00:31:05 -	00:01:41 x 18	00:01:42 x 18	00:04:20 x 7
Stop times Benefit against standard	00:01:38 -	00:01:00 X 1,6	00:01:01 x 1,6	00:01:32 x 1,1

Figure 3-22 Start and stop times and their benefits

The time values are based on HH:MM:SS. The start and stop time information's are extracted from the index server trace files of the SAP HANA environment. All tests are running up to minimum 3 times.

The result values that are listed in Figure 3-22 are average values. It is important to emphasize that the factors that the start improved do not correlate to the run time and heavily depend on the amount of data that is inside HANA. Therefore, it is recommended to validate whether the targeted business application supports SAP HANA Native Storage Extension (NSE) to first start reducing the amount of memory that is loaded and the overall memory consumption.

Based on older measurements when moving to an FS9100 model, a start time improvement of a minimum factor of 2 - 3x can be expected. For the PCIe NVMe local disks, 5x start improvements can be easily achieved.

3.6.3 Mapping to H922 and H924 models

In 2019, IBM released new H922 and H924 bundles that were documented in the IBM Power Systems H922 and H924 Technical Overview and Introduction publication. Depending on the size, these releases include 3 or 5 NVMe cards that are preconfigured and provide protection from a single card failure (if higher protection is required, adding one card provides protection from two card failures).

3.7 Effect on Live Partition Mobility capabilities

PowerVM LPM technology allows for the migration of LPARs from one system to another (offline mode or live). This technology allows LPARs to be moved to another system to avoid downtime to the running applications.

The use of tmpfs for persistent memory uses the memory that is assigned to the LPAR and available to the LPAR. Therefore, moving an LPAR from one system to another preserves the use of the tmpfs persistent memory volumes at the destination system.

vPMEM volumes that are assigned to an LPAR are defined outside the LPAR configuration. Because of this implementation, LPM operations are not supported for vPMEM-enabled LPARs.

Before the LPAR migration, the vPMEM device must be removed from the LPAR. Then, a new vPMEM volume can be created at the destination system to support the application. vPMEM LPM is intended to be supported in a future firmware release.

4

SAP HANA memory footprint effects on native storage expansion

This chapter describes the memory effects of native storage expansion (NSE) on SAP HANA.

This chapter includes the following topics:

4.1 SAP HANA NSE overview

SAP HANA Native Storage Extensions (NSE) is a new feature of SAP HANA starting with SAP HANA 2 SPS 04. This SAP HANA feature enables dropping cold or warm data objects from the Column Store to disk and handle them in a separate buffer area that is called BUFFER CACHE if needed. All data objects (Column Store and NSE) are stored in `/hana/data`. Only the load time of the objects differs.

When the business applications need access to warm data that is NSE enabled, they are found directly in the BUFFER CACHE or are loaded only into the BUFFER CACHE. For improved optimization of BUFFER CACHE usage, only needed parts of the data objects are loaded into this memory area.

Figure 4-1 shows the technical design of the NSE implementation.

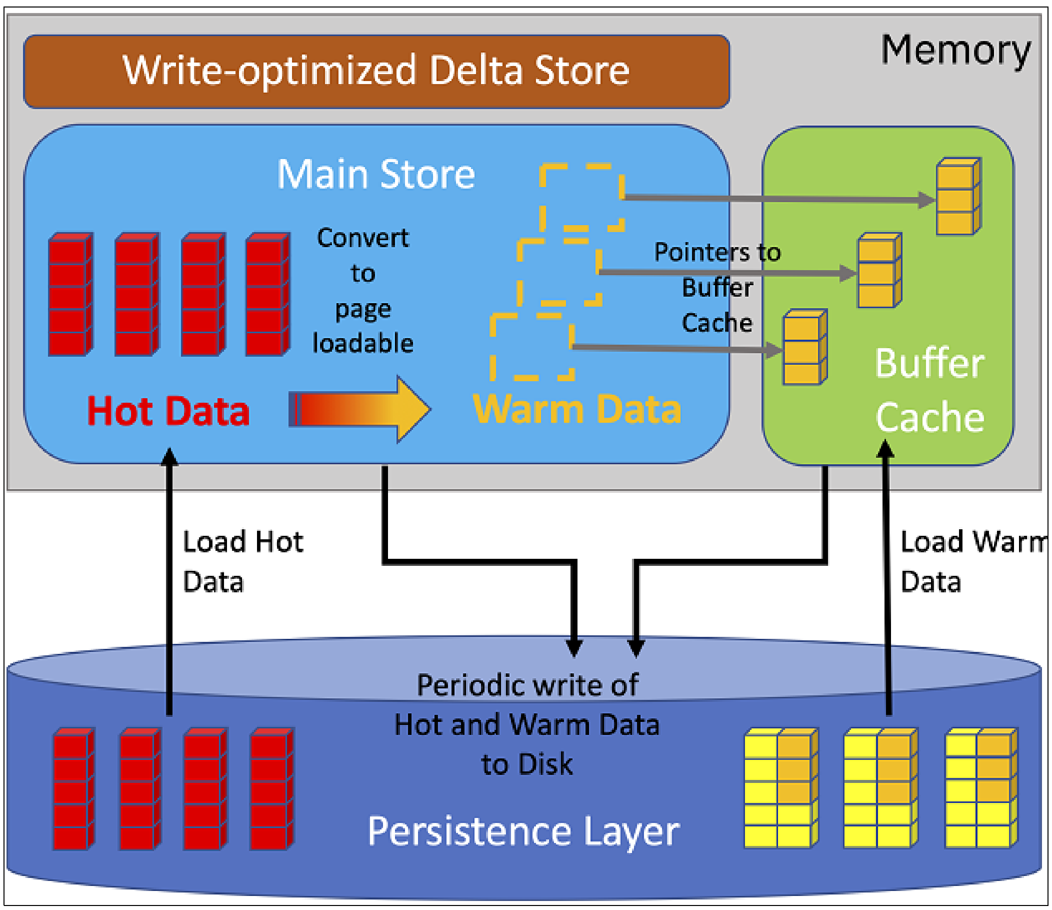


Figure 4-1 Technical NSE Design overview

Note: The SAP HANA NSE feature has no effect on IBM Power feature Live Partition Mobility if the LPAR includes no dedicated resources.

SAP HANA NSE can reduce the memory footprint on IBM Power Systems servers depending on data and workload by:

- Decreasing the Global Allocation Limit of the SAP HANA instances
- Minimizing the need of adapting SAP HANA Extension Nodes

All SAP HANA NSE features can be used and changed in real time, without direct effects on the application availability.

4.1.1 SAP HANA buffer cache and buffer cache pools

By default, the buffer cache uses 10% of the SAP HANA memory (Global Allocation Limit) but can be adjusted. The buffer cache normally is enabled, but can be disabled and reenabled.

Two SAP HANA views are available for monitoring. The first view is the M_BUFFER_CACHE_STATISTICS system view is shown in Example 4-1, which provides information about the buffer cache configuration, buffer cache state, and the memory usage.

Example 4-1 Example of view M_BUFFER_CACHE_STATISTICS

```
|HOST|PORT|VOLUME_ID|CACHE_NAME|STATE|REPLACEMENT_POLICY|MAX_SIZE|ALLOCATED_SIZE|U
SED_SIZE|BUFFER_REUSE_COUNT|HIT_RATIO|
|"lsh30041"|30003|3|"CS"|"ENABLED"|"IMPROVED
LRU"|425554083840|425553926400|413086376448|1855971|98.4456100463867|
|"lsh30041"|30007|2|"CS"|"ENABLED"|"IMPROVED LRU"|425554083840|0|0|0|0|
```

In addition, this monitoring view provides information about the quality of the buffer cache, such as hit ratio and reuse count.

The second view is the M_BUFFER_CACHE_POOL_STATISTICS system view, as shown in Example 4-2, which provides statistics about each pool in the buffer cache. The buffer cache is split in multiple pools. Each pool includes a specific page size 4 KB - 16 MB. The ALLOCATE_SIZE is the amount of all BUFFER_SIZE and TOTAL_BUFFER_COUNT products:

```
ALLOCATE_SIZE = SUM(BUFFER_SIZE * TOTAL_BUFFER_COUNT)
```

Note: The HIT_RATIO value of the BUFFER_CACHE_STATISTICS view is nearly 98% in production systems (less for non-production is possible depending on performance needs).

Example 4-2 Example of view M_BUFFER_CACHE_POOL_STATISTICS

```
|HOST|PORT|VOLUME_ID|CACHE_NAME|BUFFER_SIZE|REPLACEMENT_POLICY|GROWTH_PERCENT|TOTA
L_BUFFER_COUNT|FREE_BUFFER_COUNT|LRU_LIST_BUFFER_COUNT|HOT_BUFFER_COUNT|BUFFER_REU
SE_COUNT|OUT_OF_BUFFER_COUNT|
|"lsh30041"|30003|3|"CS"|4096|"IMPROVED LRU"|1|852472|851999|382|91|0|0|
|"lsh30041"|30003|3|"CS"|16384|"IMPROVED LRU"|1|246269|245983|162|124|0|0|
|"lsh30041"|30003|3|"CS"|65536|"IMPROVED LRU"|1|64058|59647|171|4240|0|0|
|"lsh30041"|30003|3|"CS"|262144|"IMPROVED LRU"|8|1569295|5|3036|1566254|1855971|0|
```

4.1.2 SAP HANA load units

SAP HANA NSE provides the following types of possible granularities to enable data objects to use NSE BUFFER CACHE to enable:

- The complete column store table to use the NSE BUFFER CACHE
- One or more partitions of a table to use NSE BUFFER CACHE
- Only specific columns of a column store table for BUFFER CACHE

For more information and instructions about enabling or disabling data objects, see the SAP HANA Native Storage Extension Help Page.

4.1.3 Other useful monitoring views

The following system views are helpful to observe and check relevant values of column store objects that are in the NSE contents:

- M_CS_TABLES, M_CS_COLUMNS contains the LOAD_UNIT
- M_TABLE_PERSISTENCE_STATISTICS includes the DISK_SIZE

4.2 SAP HANA NSE Advisor

The SAP HANA NSE Advisor is a feature of SAP HANA that is used to identify data objects for enabling NSE or returning NSE-enabled data objects to Column Store, if the usage is too high or new workloads are introduced (both as online operations).

The Advisor is disabled by default and must be enabled by using an SQL command or SAP HANA Studio. Before enabling the SAP HANA NSE Advisor, it is recommended to clean the Access State Cache.

To get useful results, it is recommended to run the NSE Advisor over several hours, where a representative workload is run. The following commands enable or disable the SAP HANA NSE Advisor and clean the Access Stats Cache:

- ALTER SYSTEM CLEAR CACHE ('cs_access_statistics')
- ALTER SYSTEM ALTER CONFIGURATION ('indexserver.ini','system') SET('cs_access_statistics','collection_enabled') = 'true' WITH RECONFIGURE
- ALTER SYSTEM ALTER CONFIGURATION ('indexserver.ini','system') SET ('cs_access_statistics','collection_enabled') = 'false' WITH RECONFIGURE

4.2.1 Workflow

To make the SAP HANA database ready for the optimal use of the NSE feature, a workflow is available for the NSE Advisor use, which can be followed to perform the correct NSE setup for recommended data objects.

The SAP HANA NSE Advisor generates more load on the system; therefore, it is recommended to enable the NSE Advisor for analyzing and verifying the NSE table and system settings only.

4.2.2 SAP HANA NSE Advisor recommendations

The SAP HANA NSE Advisor adds the recommended information into a separate SYSTEM view. This view is named SYS.M_CS_NSE_ADVISOR and can be accessed by way of an SQL select statement as shown in Example 4-3. One row is in this view per data object, depending on the granularity. This SQL select statement contains the following information:

- HOST, PORT, SCHEMA_NAME, TABLE_NAME

  The main information about the identified data object.

- COLUMN_NAME

  Shows the COLUMN NAME, if the data object is only a column of the identified table.

- PART_ID

  Displays the PARTITION ID, if only a specific partition of a table is identified.

- LOAD_UNIT

  The LOAD UNIT can be COLUMN or PAGE, depending on the recommendation to enable this data object for NSE or disable it.

- GRANULARITY

  Shows the granularity of the data object, which can be TABLE, PARTITION, or COLUMN.

- MEMORY_SIZE_IN_MAIN

  Views the size in memory of the recommended data object in bytes.

- MAIN_PHYSICAL_SIZE

  Displays the storage size of the data object in bytes.

Example 4-3 SQL select statement to view SYS.M_CS_NSE_ADVISOR

```
> select TABLE_NAME,COLUMN_NAME,PART_ID,LOAD_UNIT,GRANULARITY,
MEMORY_SIZE_IN_MAIN,MAIN_PHYSICAL_SIZE from SYS.M_CS_NSE_ADVISOR
TABLE_NAME,COLUMN_NAME,PART_ID,LOAD_UNIT,GRANULARITY,MEMORY_SIZE_IN_MAIN,MAIN_PHYS
ICAL_SIZE
"/BA7/PMAT_PLANT","NULL",0,"PAGE","TABLE",1423850960,398528512
"/BA7/SMATL","NULL",0,"PAGE","TABLE",16292991,14278656
"/BA7/XMATL","NULL",0,"PAGE","TABLE",37928343,36429824
"/BA7/XMAT_PLANT","NULL",0,"PAGE","TABLE",1990987164,398528512
"/BA7/AB4REPA23","/BA7/S_CURKYF01",0,"PAGE","COLUMN",754442296,790777856
"/BA7/AB4REPA23","/BA7/S_UOMKYF01",0,"PAGE","COLUMN ",588361336,61107404
```

4.2.3 SAP HANA NSE Advisor configurations

The SAP HANA NSE Advisor uses specific rule-based statistics to identify the temperature of the data objects. This kind of heuristics can be modified by the following NSE Advisor parameters:

- The Scan density influences the data object access rate and memory ratio.
- The Hot object threshold is a percentage value that influences the minimum scan density for a data object to be a considered a hot object. The default is 10%.
- The Cold object threshold is a percentage value that influences the maximum scan density for a data object to be a considered a cold object. The default is 10%.
- The Minimum object size is the minimum size of the data objects that are considered for NSE Advisor recommendations. This threshold default is 1 MB.

Figure 4-2 shows the effect of the NSE Advisor configuration in terms of expected recommendations.

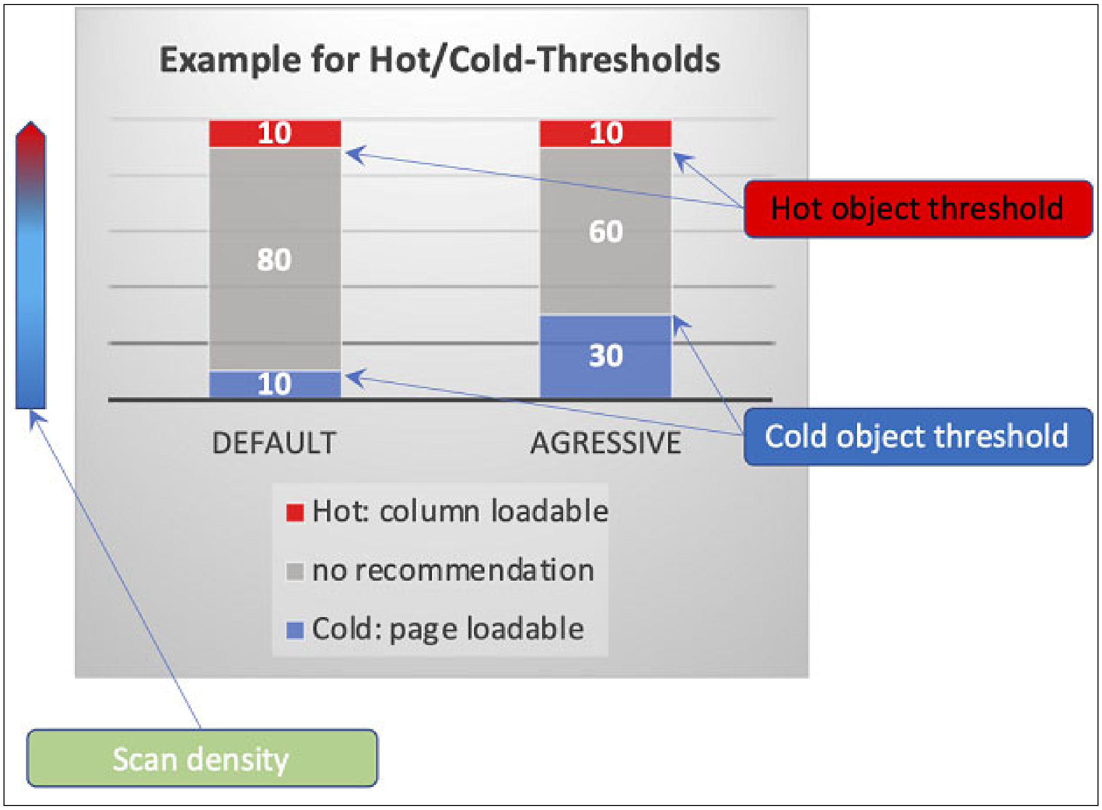


Figure 4-2 Configuration effect on NSE Advisor work

4.3 Performance and memory effects by using different NSE setups

This section describes how different NSE configurations affect the application performance and memory footprint of the SAP HANA database system. A business warehouse system was used. This system is based on SAP BW/4 HANA. The SAP HANA database server runs SAP HANA 2 SPS 05.

> **Note:** These results depend on a specific SAP HANA BW/4 system, system setup, and workload. Results on other SAP solutions that are based on SAP HANA 2.0 can show a different picture. These results are based on application workload, the hardware infrastructure, total persistence size, and the relation between cold, warm, and hot data.
>
> The improvements that are described next cannot be used as reference savings.

The following scenarios were observed:

- Standard configuration of SAP HANA:
  - NSE is enabled, but no data object is configured as PAGE usage.
  - All Data reside in the Column Store with its persistence on disk.
- The biggest business warehouse table, which is recommended for NSE use, is switched to PAGE mode:
  - This table takes nearly 25% of the complete database.
  - During this configuration change, the buffer cache uses and influences the overall memory usage of SAP HANA. The real memory savings are less than 25%.
- All business warehouse tables, which are recommended by the application owner, are converted to PAGE mode:
  - These tables are covering 70% of the hole database data footprint.
  - The real memory savings are less than 70% because the buffer cache is fully used here.
- All business warehouse tables, including all hot tables, are switched to buffer cache use. This configuration is added to highlight the value of faster Storage back ends to absorb some performance effects:
  - This configuration enables 99% of all tables of the complete database to PAGE mode.
  - This test scenario shows how stable the underlying storage infrastructure on POWER operates.

To get test results that are as realistic as possible, the SAP HANA BW/4 system was tested with two different types of query workloads. To simulate business workload of users', complex queries are used. To measurement the effect on batch workload, massive parallel simple queries used. Consider the following points:

- The simple query test runs a set of different simple queries in parallel over a specific time. The number of queries is measured over this specific time. The number of parallel queries that are run is configured such that the overall CPU use increases to nearly 90% usage.
- The complex queries are run as single query and the measurement is based on the query run time.

4.4 Memory savings by using NSE-enabled data objects

The effect on the main memory usage of SAP HANA can be reduced significantly by using the SAP HANA NSE feature.

Figure 4-3 shows the memory savings in the test SAP HANA BW/4 environment.

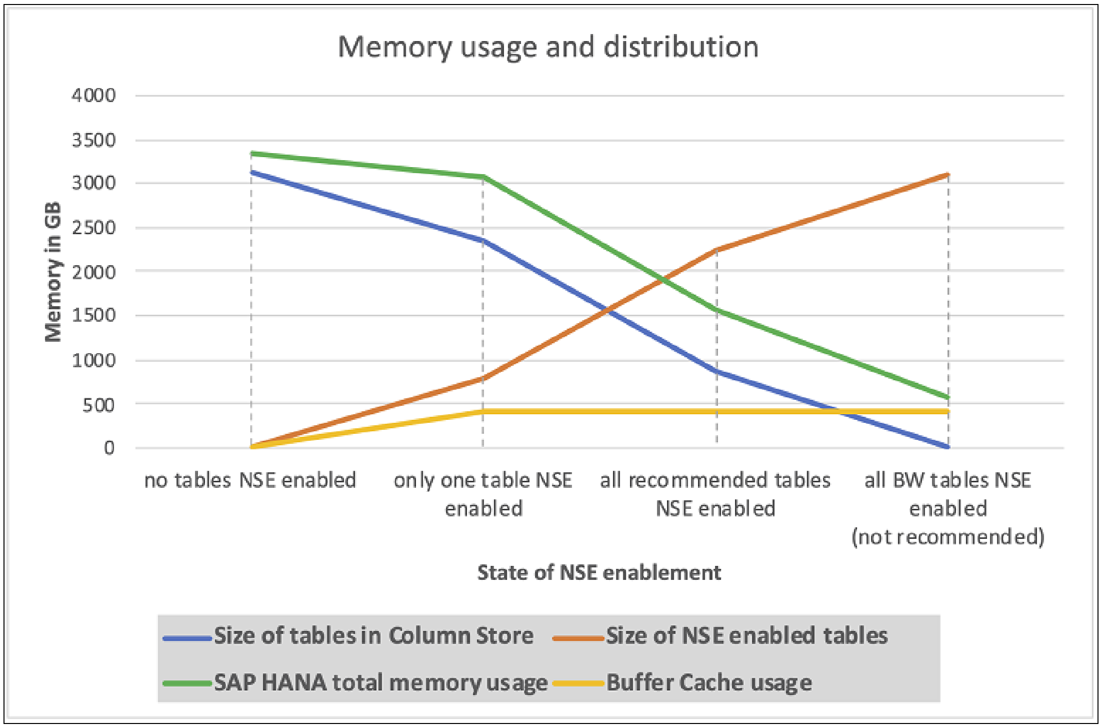


Figure 4-3 SAP HANA Memory usage

The total memory usage in a standard setup without NSE enabled tables of the test setup amounts to nearly 3.4 TB of main memory. If all recommended tables are switched to NSE to use the buffer cache, the overall main memory usage is under 1.6 TB. To have stable and transparent test results, the buffer cache size was not changed during the test executions. The buffer cache is 400 GB in these test scenarios.

4.5 Effect on application performance

The use of the SAP HANA NSE feature for recommended cold and warm data has no or only minimal effect on the application performance. Figure 4-4 shows the performance effect on different NSE configurations.

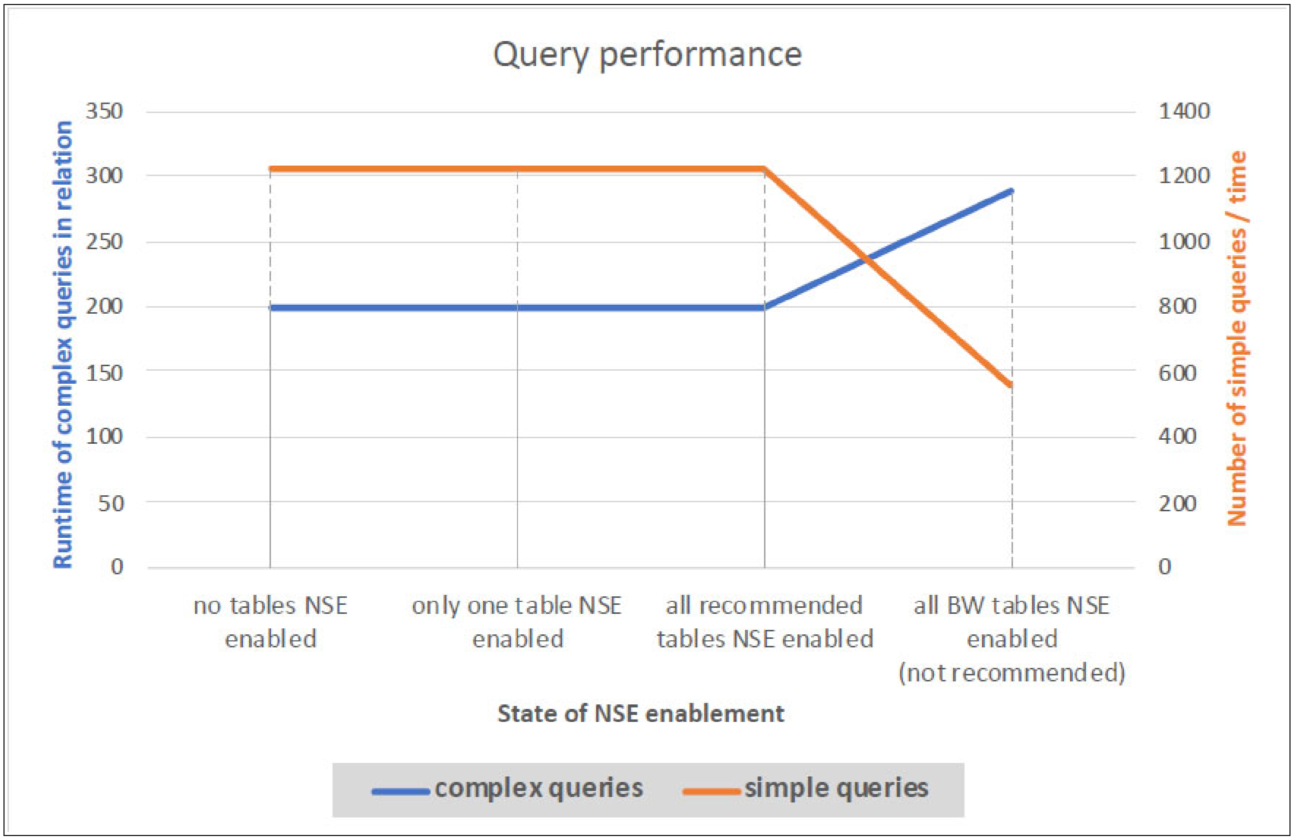


Figure 4-4 Query performance on different NSE configurations

Only when a non-recommended NSE configuration is used (where hot data is enabled for NSE) does the application performance goes down for simple queries. In some cases, if the storage environment of the SAP HANA persistence is too slow, the application can suffer significant performance deterioration.

Figure 4-4 shows the results of a test that was performed on a fast storage solution that was configured with direct attached NVMe devices. The buffer cache was limited to 400 GB. If the buffer cache size is increased, the performance affects only the worst-case scenario. Therefore, increasing the SAP HANA buffer cache help to reduce the performance degradation of the application.

4.6 I/O impact by using not recommended NSE configuration

The I/O effect on the non-recommended use of the SAP HANA NSE feature can be dramatic. In the case of a NSE-enabled data object that is switched from cold or warm to hot, the I/O traffic to the persistence memory (storage layer) increases. This issue can occur when an application changes or massive access to data objects occurs that normally are cold or warm, such as historical data.

To protect the ability of work for the SAP HANA system, a fast storage subsystem can be helpful. Figure 4-5 shows the performance effects on different storage subsystem solutions for the SAP HANA persistence layer.

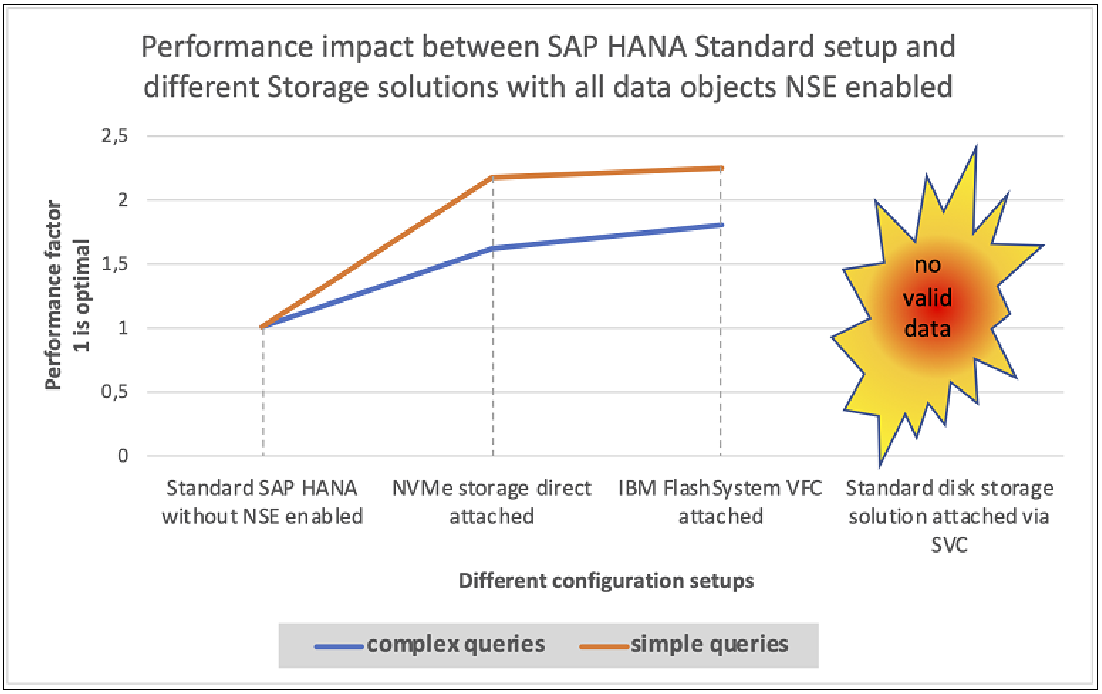


Figure 4-5 I/O performance impact on not recommended NSE configurations

4.7 Start times by using NSE enabled data objects

In addition to the memory savings that are achieved by using the SAP HANA NSE feature, the SAP HANA start time also decreases. This start time decrease has a positive effect on planned and unplanned downtimes.

Figure 4-6 shows the start time of an SAP HANA BW/4 system with a persistence size of 3.2 TB. In this case, a reduction of factor 4 between standard SAP HANA configuration and SAP HANA with recommended NSE enabling was possible.

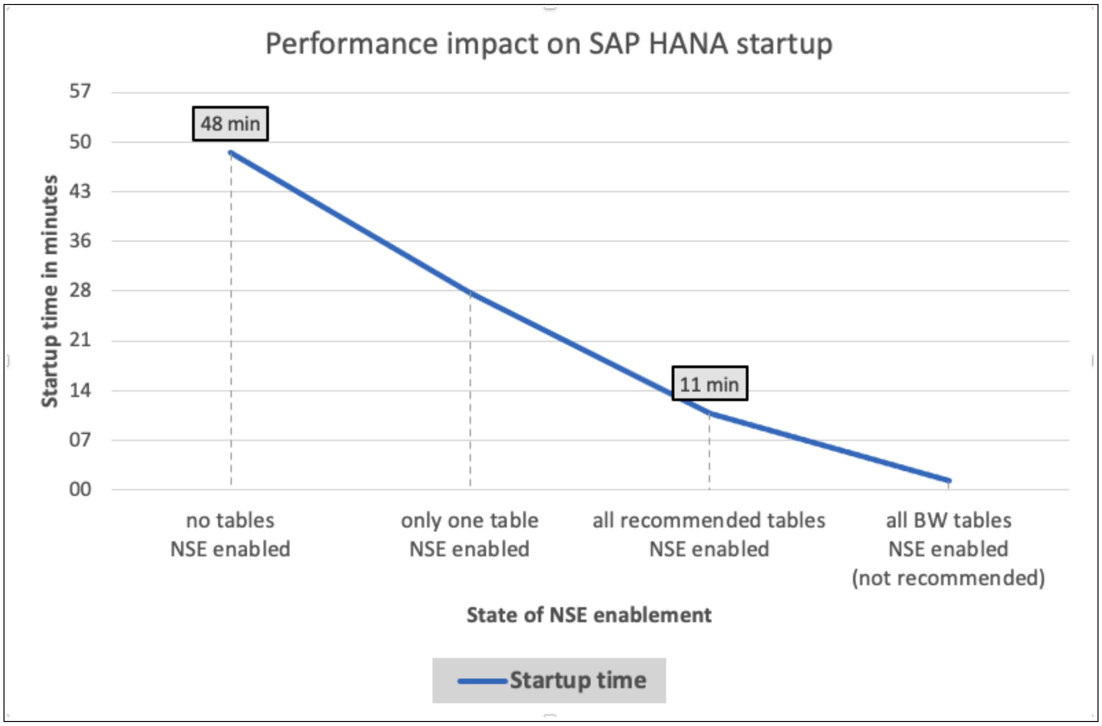


Figure 4-6 SAP HANA start times on different NSE scenarios

Related publications

The publications that are listed in this section are considered suitable for a more detailed description of the topics that are covered in this paper.

IBM Redbooks

The following IBM Redbooks publications provide more information about the topics in this document. Some publications that are referenced in this list might be available in softcopy only:

- *IBM Power Systems Security for SAP Applications*, REDP-5578
- *IBM Power Systems Virtualization Operation Management for SAP Applications*, REDP-5579
- *SAP HANA on IBM Power Systems: High Availability and Disaster Recovery Implementation Updates*, SG24-8432
- *SAP Landscape Management 3.0 and IBM Power Systems Servers*, REDP-5568

You can search for, view, download, or order these documents and other Redbooks, Redpapers, web docs, drafts, and other materials, at the following website:

ibm.com/redbooks

Online resources

The following websites are also relevant as further information sources:

- Guide Finder for SAP NetWeaver and ABAP Platform:

  https://help.sap.com/viewer/nwguidefinder
- IBM Power Systems rapid cold start for SAP HANA:

  https://www.ibm.com/downloads/cas/WQDZWBYJ
- SAP Support Portal:

  https://support.sap.com/en/index.html
- Software Logistics Tools:

  https://support.sap.com/en/tools/software-logistics-tools.html
- Welcome to the SAP Help Portal:

  https://help.sap.com

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

REDP-5570-01

ISBN 0738459860

Printed in U.S.A.