SAP HANA ™
Storage Requirements

As an in-memory database, SAP HANA uses storage devices to save a copy of the data, for the purpose of startup and fault recovery without data loss. The choice of the specific storage technology is driven by various requirements like size, performance and high availability. This paper discusses the SAP HANA storage requirements.

SAP HANA Development Team

V3.0, March 2024

Contents

Change history

Version	Date	Description
1.9	June 2013	Initial release
2.0	November 2013	Added Storage Sizing section Added Performance Section
2.1	December 2013	Improved information about IO fencing with NFS based shared storages
2.2	May 2014	Updated sizing formulas Minor fixes
2.3	July 2014	Minor fixes
2.4	October 2014	Updated Sizing chapter: New formula for redo log sizing
2.5	January 2015	Updated Sizing chapter: Distinguish total memory and net data size on disk
2.6	March 2015	Updated Sizing chapter: New recommendations for sizing of /hana/shared
2.7	February 2016	Added chapter: Physical separation of data and log volumes
2.8	December 2016	Reworked Sizing chapter
2.9	January 2017	Updated Sizing chapter: Double disk space for data during table redistribution
2.10	February 2017	Updated storage sizing chapter: (1) New note in “preface” section (2) Improved wording for data volume formulas
3.00	March 2024	Updated sizing data & minor corrections

Introduction

SAP HANA is an in-memory database which stores and processes the bulk of its data in memory. Additionally, it provides protection against data loss by saving the data in persistent storage locations. For setting up an SAP HANA system, the storage layer must fulfill several requirements. This paper discusses the different requirements and common design options for the storage subsystem. Especially when using high availability and disaster tolerance features, care must be taken on planning the persistent space.

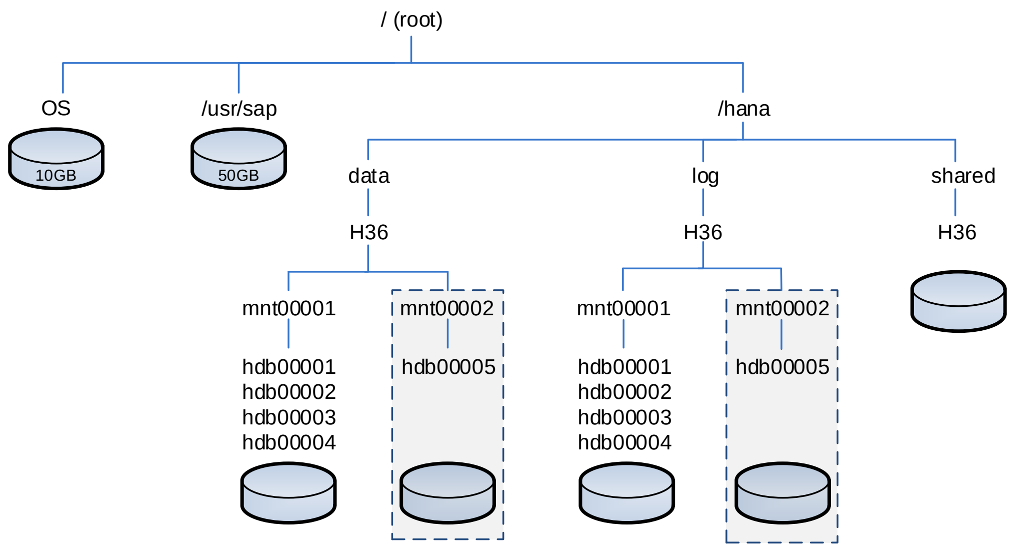
SAP HANA uses storage for several purposes:

- The SAP HANA *installation.* This directory tree contains the run-time binaries, installation scripts and other support scripts. In addition, this directory tree contains the SAP HANA configuration files*,* and is also the default location for storing trace files and profiles. On distributed systems, it is created on each of the hosts.
- *Backups*. Regularly scheduled backups are written to storage in configurable block sizes up to 64 MB.
- *Data*. SAP HANA persists a copy of the in-memory data, by writing *changed* data in the form of so-called Savepoint blocks to free file positions, using I/O operations from 4 KB to 16 MB (up to 64 MB when considering super blocks) depending on the data usage type and number of free blocks. Each SAP HANA service (process) separately writes to its own Savepoint files, every five minutes by default.
- *Redo Log*. To ensure the recovery of the database with zero data loss in case of faults, SAP HANA records each transaction in the form of a so-called redo log entry. Each SAP HANA service separately writes its own redo-log files. Typical block-write sizes range from 4KB to 1MB.

Each service of a distributed (multi-host) SAP HANA system manages its persistence independently. Logically, this is a *shared-nothing* approach.

Conceptual Storage Layout

The figure below depicts the recommended file system structure of an SAP HANA system (named “H36” in this example) which is distributed over several hosts (two, in this example).



The /usr/sap location is the place where additional SAP software required by SAP HANA will be installed; SAP recommends reserving 50 GB of disk space for it. Moreover, SAP recommends reserving another 10 GB of disk size to hold the plain OS installation. The location of the OS installation can be joined with the /usr/sap location.

Unless specified differently at installation time, the installation binaries will reside at the /hana/shared/<sid> location; <sid> referring to the 3-digit system ID. The data files will reside at /hana/data/<sid>, the log files at /hana/log/<sid>.

To answer the question what kind of storage device(s) to use for all those file system locations, more aspects need to be considered. For a single-host SAP HANA system that does not require any measures for failure safety, one might simply use local storage devices delivered as integral part of the compute server[1], such as SCSI hard drives or SSDs, for all file system locations depicted above. However, this may change as soon as a certain degree of failure safety and protection against data loss needs to be guaranteed, examples:

- For regular backups of a distributed SAP HANA system, it may be desirable to store the backup archives of all instances of that system on one single storage device.
- If SAP HANA Host Auto Failover needs to be leveraged (aka "local high-availability"), it must be possible to share the same files between multiple hosts, as discussed in chapter "High Availability".

Taking such additional considerations into account, in many cases, rather than using host-internal storage, the preferred storage solution will involve separate, ***externally attached storage*** subsystem devices capable of providing dynamic mount points for the different hosts required in the overall landscape.

Regarding storage performance, both write and read operations need to be considered. A certain ***minimum speed of writing raw data to storage*** is crucial e.g. to keep latency times low when persisting redo log records in storage to complete SAP HANA database transactions.

Read speed is crucial to keep the startup times of the SAP HANA database as low as possible. The same is true in failover situations when a standby host takes over the persistence of a failing host. Data is also read during regular operations e.g. if tables are configured to be loaded into memory only if needed.

The impact of the I/O performance of the storage system on the overall behavior of an SAP HANA database is described in chapter "Performance".

Physical Separation of Data and Log Volumes

Typically, vendors of enterprise database management systems recommend a physical separation of the data and log volumes. The same is true for SAP HANA.

One reason for this recommendation is ***I/O performance***: For SAP HANA, the I/O requirements for accessing the data volume are different from those for the log volume (see chapter "I/O Sizing") so that, usually, separate file systems for data and log volume are needed to reach the I/O requirements for both volumes. The log volume is accessed using a sequential I/O pattern while the data volume is

[1] Certain performance thresholds must be accomplished.

accessed using a random pattern. Therefore, with separated volumes, a better performance can be reached for the redo log writing.

The second reason is ***data safety***: If the disks of one volume get damaged or data in one volume get corrupted then the other volume is not affected.

Storage Sizing

The formulas listed in the sections below assume that the memory sizing for the target SAP HANA database has already been done. You may refer to the following documents to learn more about it (find the URLs in the "References" chapter):

- *Sizing Approaches for SAP HANA* [1]
- *SAP Quick Sizer* [2]
- *Sizing for SAP Suite-on-HANA* [3]
- *Memory Sizing Report for SAP Suite-on-HANA* [4]
- *Sizing Report for SAP BW-on-HANA* [5]

The recommendations for storage sizing of (1) SAP HANA appliances and (2) hardware configurations according to the SAP HANA tailored data center integration (TDI) approach slightly vary. The sizing formulas listed in this document apply to TDI environments only, not to appliances. The sizing requirements for SAP HANA appliances are not discussed here.

Note

Ideally, like with any other SAP application, any sizing recommendation to meet the requirements of SAP-HANA-based applications should be based on rules specific to the application and – even more important – specific to the (expected) workload in the target environment. Put differently, any sizing for SAP HANA systems should consider the (expected) number of users and the workload those users impose on the target SAP HANA database. Therefore, note that the formulas for disk space sizing listed below in this chapter are generic rule-of-thumb formulas rather than workload-specific guidelines. While SAP believes that applying those rule-of-thumb formulas should be sufficient in terms of performance in most cases, SAP also points out that, due to their generic nature, applying the formulas below may not always result in the most cost-efficient hardware setup. Therefore, whenever more accurate application- or workload-specific guidelines for disk space sizing are available, SAP recommends applying those rather than the formulas listed below.

Memory Sizing

Earlier versions of this document used the overall amount of memory required for the target SAP HANA database (typically referred to as ***"RAM"***) as the basis for calculating the required amount of disk storage. This approach is sufficient as long as the entire database contents can be loaded into memory. However, if concepts like "hybrid LOBs" and "cold data" are leveraged in SAP HANA applications, then more sophisticated formulas must be applied.

The various SAP HANA Sizing Notes (e.g. the abovementioned programs [3] to [5]) state the result of memory sizing for a target SAP HANA database using the term ***anticipated total memory requirement***. That value includes the memory required for:

- Data stored in Row Store and Column Store tables including indexes
- Database objects created temporarily at runtime e.g. when queries are executed or when a Delta Merge operation takes place (see reference [10] for details)
- Cached records of Hybrid LOBs
    - Hybrid LOBs are stored on disk
    - It is assumed that 20% of them are cached in memory

- ~50 GB for database code, stack and other services

Note that there is a widely used ***rule-of-thumb*** to simply ***double the value of "data stored in tables including indexes"*** to come to a result value for the total memory requirement. To ease reading of the following sections in this chapter, the term ***"RAM"*** will be used ***as synonym for "anticipated total memory requirement"***.

Disk Space Required for the Data Volume

Whenever a Savepoint or a Snapshot is created or a Delta Merge operation is performed, data is written from memory to the data volume located at the mount point `/hana/data/<sid>`.

Option 1: ***If an application-specific sizing program can be used*** (e.g. the SAP HANA Quick Sizer [2] or one of the sizing programs [3], [4] or [5]), ***the recommended size of the data volume*** of a given SAP HANA system ***is equal to the value of the "anticipated net data size on disk"*** (as stated in the result of the sizing program) ***plus an additional free space of 50%***:

$$\textbf{Size}_{\textbf{data}} \textbf{ = 1.5x anticipated net data size on disk}$$

The value of "anticipated net data size on disk" consists of two parts:

- ***Net data size on disk*** = Size of all database tables including indexes
    - In an existing SAP HANA system one can retrieve this value by executing the SQL command: `SELECT SUM(disk_size)/1024/1024/1024 FROM m_table_persistence_statistics`
- ***Disk space required for merges*** = Additional disk space required during Delta Merge operations
    - During a Delta Merge the affected tables are temporarily duplicated on disk for a short period of time

Option 2: ***If there is no application-specific sizing program available*** (e.g. neither the SAP HANA Quick Sizer [2], nor one of the sizing programs [3], [4], or [5] can be applied), ***the recommended size of the data volume*** of a given SAP HANA system ***is equal to the total memory required*** for that system, according to the result of the memory sizing:

$$\textbf{Size}_{\textbf{data}} \textbf{ = 1.5 x RAM}$$

Additional Remarks

- The value for the disk space of the data volumes calculated using one of the two formulas above refers to the entire SAP HANA system. That is, if the SAP HANA system is distributed over several worker nodes with each worker node having its own data volume, then the size of the data volume for one worker must be calculated by dividing the overall value by the number of worker nodes of that system.

- During the migration of a non-SAP-HANA database to SAP HANA, the system may temporarily need more disk space for data than calculated in the sizing phase. This is not reflected in the formula above. It is assumed that the storage system can provide the SAP HANA system with that additional space if required.

- During a table redistribution process, the system needs temporarily more disk space for data than calculated in the sizing phase. As a rule-of-thumb, the size of the disk space for data should be doubled during redistribution to avoid running in disk-full situations.

- Make sure that for the sizing of the data volume you consider data growth and fragmentation of the data volume. Regarding the latter, the SQL command `ALTER SYSTEM RECLAIM DATAVOLUME` can be used to optimize the size of the data volume manually – see the SAP HANA documentation for more details. Starting from SAP HANA 2 SPS06 an automatic data volume reclaim ensures that not too much fragmentation exists.

- In high availability and disaster recovery setups which leverage either SAP HANA System Replication or storage replication, the same amount of storage must be provided on both primary and secondary sites.

- In case of high usage of Hybrid LOBs, SAP HANA Native Storage Extension or having long running data backup operations or long running transactions, the required data volume size might be even higher. For that reason, close monitoring of the data volume usage and its free size is important to avoid interruptions.

Disk Space Required for the Log Volume

The minimum size of the log volume depends on the number of data changes occurring between two SAP HANA Savepoints which – per default – are created every 5 minutes. The more data changes are executed by write transactions in that period of time, the more redo log segments are written to the log volume under `/hana/log/<sid>`. When sizing the log volume, the following points must be considered:

- The redo log must not be overwritten before a Savepoint entry is available in the data volume; otherwise, the SAP HANA database may become unable to restart.

  Situations may occur where the writing of a Savepoint is delayed, for example if very high workload needs to be processed during a database migration process in an environment with rather slow I/O between source and target (SAP HANA) database. In such cases, as long as the Savepoint has not been written to the data volume, the amount of redo logs in the log volume will keep on growing until all log segments are full.

- If “log_mode = normal” is set the redo log must not be overwritten before a backup took place. Therefore, it is recommended to have some extra space available for situations where incidents or faults may interrupt the backup process. That extra space should allow for system administrators to fix and finish the backup process before the log volume runs full.

There is no direct correlation between the SAP HANA database size and the required log volume size. Nevertheless, we recommend using the formula below as a rule of thumb since it is based on best practice and experiences with productive SAP HANA installations. The value depends on the total memory requirement (“RAM”):

[systems ≤ 512GB] $\mathbf{Size_{redolog} = 1/2 \times RAM}$
[systems > 512GB] $\mathbf{Size_{redolog(min)} = 512GB}$

Examples:

- 128 GB system => $Size_{redolog}$ = 64 GB
- 256 GB system => $Size_{redolog}$ = 128 GB
- 512 GB system => $Size_{redolog}$ = 256 GB
- 1 TB system => $Size_{redolog(min)}$ = 512 GB
- 2 TB system => $Size_{redolog(min)}$ = 512 GB
- 4 TB system => $Size_{redolog(min)}$ = 512 GB
- …

Additional Remarks

- The value for the disk space of the log volumes calculated using the formulas above refers to the entire SAP HANA system. That is, if the SAP HANA system is distributed over several worker nodes with each worker node having its own log volume, then the size of the log volume for one worker must be calculated by dividing the overall value by the number of worker nodes of that system.
- For systems with more than 512 GB in-memory database size, the formula above represents a minimum value. As of today, based on the experience made with productive SAP-internal SAP HANA installations, this value is considered sufficient for each SAP HANA use case. Nevertheless, as described above, as the amount of data stored in the log volume depends on the workload processed, there may be situations where this value is not sufficient for log volume sizing.
- In high availability and disaster recovery setups which leverage either SAP HANA System Replication or storage replication, the same amount of storage must be provided on both primary and secondary sites.

Disk Space Required for SAP HANA Installation

All binary, trace and configuration files are stored on a shared file system that is exposed to all hosts of a system under `/hana/shared/<sid>`. Thus, additional space is required for the traces written by the compute node(s) of the SAP HANA database. Experiences with productive SAP HANA installations show that the bigger the size of the SAP HANA database, the more traces are written. Therefore, the calculation is based on the total memory requirement (“RAM”).

For **single-node** SAP HANA systems, the recommended disk space for `/hana/shared/<sid>` is:

$$Size_{installation(single\text{-}node)} = MIN(1 \times RAM; 1\ TB)$$

Examples:

- Single-node 128 GB => $Size_{installation}$ = 128 GB
- Single-node 256 GB => $Size_{installation}$ = 256 GB
- Single-node 512 GB => $Size_{installation}$ = 512 GB
- Single-node 1 TB => $Size_{installation}$ = 1 TB
- Single-node 2 TB => $Size_{installation}$ = 1 TB
- Single-node 4 TB => $Size_{installation}$ = 1 TB
- Single-node 6 TB => $Size_{installation}$ = 1 TB

For **scale-out** SAP HANA systems, the recommended disk space for `/hana/shared/<sid>` depends on the number of worker nodes. Per each four worker nodes of a given scale-out system, a disk space of 1x RAM of one worker is recommended:

$Size_{installation(scale\text{-}out)}$ = 1 x RAM_of_worker per 4 worker nodes

Examples:

- 3+1 system, 512 GB per node => $Size_{installation}$ = 1x 512 GB = 512 GB
- 4+1 system, 512 GB per node => $Size_{installation}$ = 1x 512 GB = 512 GB
- 5+1 system, 512 GB per node => $Size_{installation}$ = 2x 512 GB = 1 TB
- 6+1 system, 512 GB per node => $Size_{installation}$ = 2x 512 GB = 1 TB
- 7+1 system, 512 GB per node => $Size_{installation}$ = 2x 512 GB = 1 TB
- 8+1 system, 512 GB per node => $Size_{installation}$ = 2x 512 GB = 1 TB
- 9+1 system, 512 GB per node => $Size_{installation}$ = 3x 512 GB = 1.5 TB
- …
- 3+1 system, 1 TB per node => $Size_{installation}$ = 1x 1 TB = 1 TB
- 4+1 system, 1 TB per node => $Size_{installation}$ = 1x 1 TB = 1 TB
- 5+1 system, 1 TB per node => $Size_{installation}$ = 2x 1 TB = 2 TB
- 9+1 system, 1 TB per node => $Size_{installation}$ = 3x 1 TB = 3 TB
- …

Disk Space Required for Backups

A complete data backup contains the entire payload of all data volumes. The size required by the backup directory not only depends on the total size of the data volumes, but also on the number of backup generations kept on disk and on the frequency with which data is changed in the SAP HANA database. For example, if the backup policy requires to perform complete data backups daily and to keep those backups for one week, the size of the backup storage must be seven times the size of the data area.

In addition to data backups, backup storage for log backups must be reserved to provide the possibility for a point-in-time database recovery. The number and size of log backups to be written depend on the number of change operations in the SAP HANA database.

$Size_{backups} \geq Size_{data} + Size_{redolog}$

Technically, it is possible to store the backups of several SAP HANA databases in a central shared backup storage. But if several backup and/or recovery processes run in parallel, this will have an impact on the overall data throughput of the given backup storage. That is, backup and recovery processes may slow down significantly, if the backup storage cannot guarantee a constant level of data throughput once the number of parallel processes exceeds a certain number.

Disk Space Required for Exports

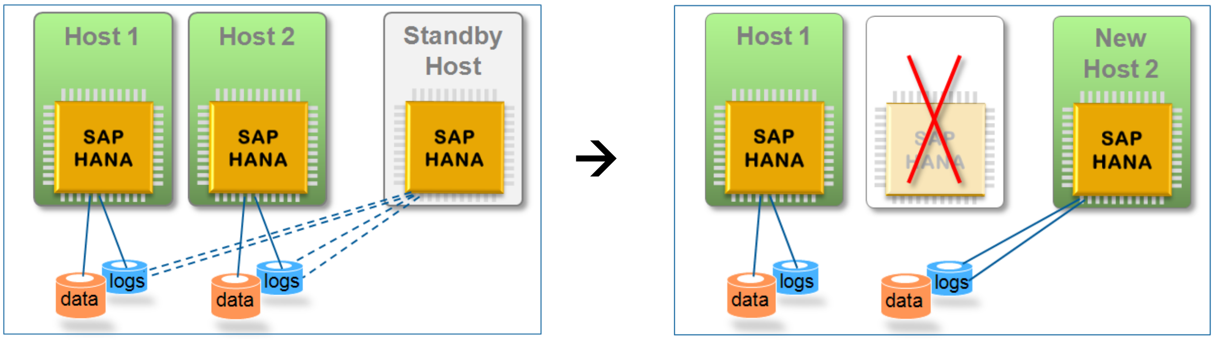
Sometimes the database content is needed for a root cause analysis of problems. For this purpose, sufficient disk space must be provided to hold the binary exports. In most cases it is not necessary to export the entire database content for root cause analysis. Therefore, as a rule of thumb it should be sufficient to reserve storage space of about two times the size of the biggest database table.

High Availability

The *SAP HANA High Availability White Paper* [6] describes the different SAP HANA High Availability (HA) solutions offered to support rapid recovery from faults and disasters. Each of these solutions represents a tradeoff and yields different storage requirements.

Failure Recovery: Host Auto-Failover

The following illustration shows how Host Auto-Failover logically works. An active host fails (e.g. Host 2), and the Standby host takes over its role by starting its database instance using the persisted data and log files of the failed host (Host 2).



For this scenario to work, three conditions must exist:

1. Failures must be detected reliably, and a reliable failover must follow, such that the standby host takes over the role
2. Upon failover, the database on the standby host must evidently have read and write access to the files of the failed active host.
3. Upon failover, the original, failed host must no longer be able to write to these files, or else these files might become corrupted. Ensuring this is called *fencing*.

Failure Detection and Failover

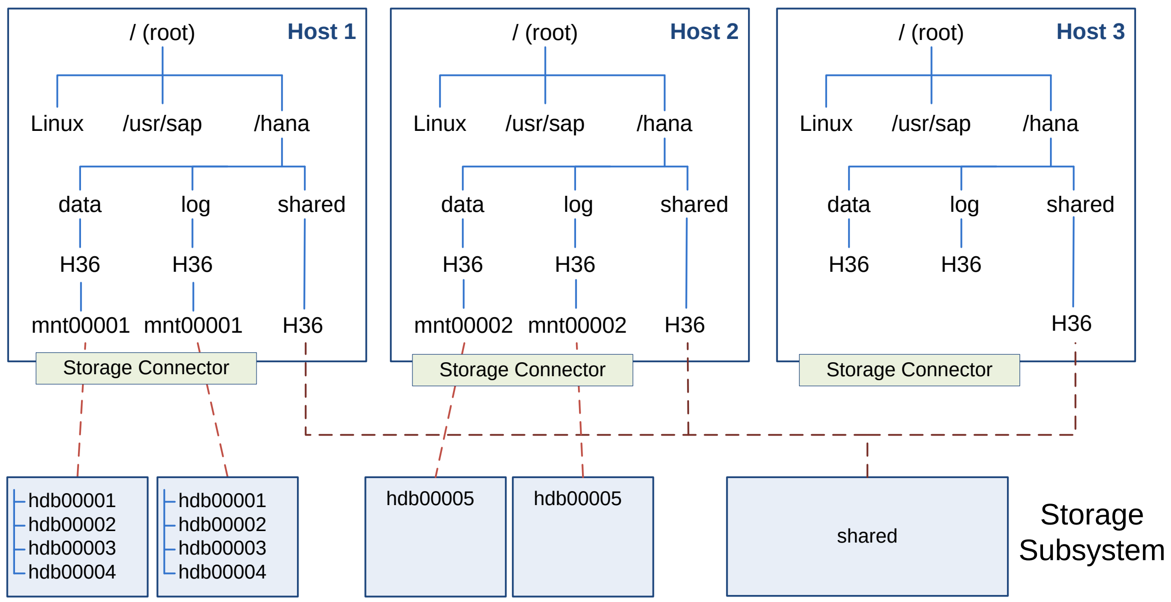
The failover decision is made by SAP HANA itself, making an external cluster manager unnecessary. More precisely, the name-server of one of the SAP HANA instances acts as the cluster master that pings all hosts regularly and manages the failover process and fencing requirements.

File Access and Fencing

To accomplish the file access by the standby host upon failover, as well as the fencing of the failed host, there are fundamentally two alternative design approaches. One uses separate storage devices that are re-assigned during the fail-over phase, the second design uses a shared-storage approach. These two alternatives are explained here.

Non-shared Storage

In a non-shared storage configuration, each host is connected to its own storage. Host 3 is a standby host, which has nothing mounted except for the shared area.

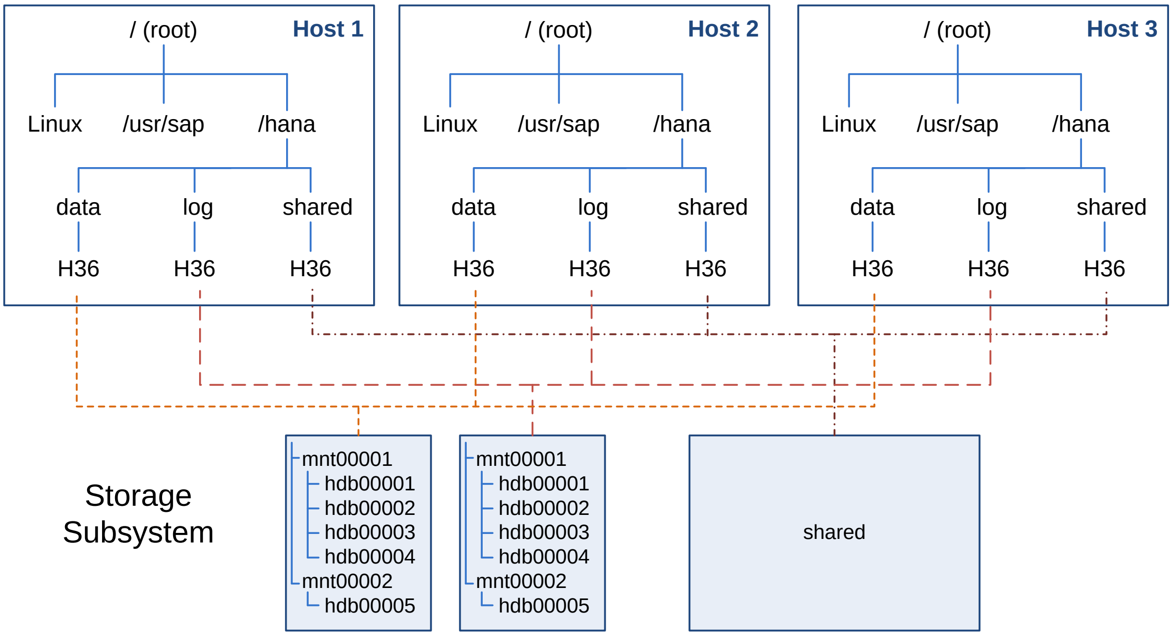


This storage architecture implies the challenge of providing a safe failover mechanism if one of the hosts fails. For this purpose, SAP HANA offers storage technology vendors an API, called the "Storage Connector API", to support file access sharing and fencing. The Storage Connector API exposes several methods, which must be implemented by the storage vendor. During failover, SAP HANA calls the appropriate Storage Connector API method, to allow the storage device driver to re-mount the required data and log volumes to the standby host and fence off these volumes from the failed host.

SAP offers a ready to use implementation of this Storage Connector API for all storage subsystems attached via Fibre Channel using native Linux multipathing and supporting the SCSI-3 protocol (SCSI-3 Persistent Reservations are used).
Speaking of Fibre Channel, note that also Fibre Channel over Ethernet (FCoE) is supported by SAP's implementation of the Storage Connector API.

Shared Storage with Shared File Systems

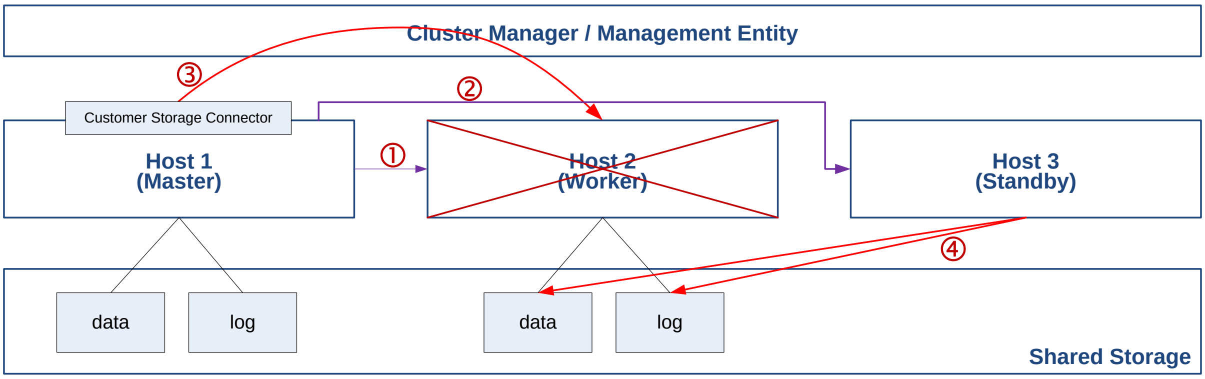
The following illustration shows three hosts using a shared storage. Host 3 is a standby host that can access all files.



Note that mounts may differ among the various hardware partners and their setups. SAP HANA requires being able to read and write all `hdb<n>` directories on all hosts regardless the actual mount points. Depending on cost, performance and maintainability considerations, data and log could also be unified to one mount point on the one hand or each `mnt<n>` directory could serve as mount point that is exposed to all hosts on the other hand. This picture shows the division of data and log that keeps the number of devices and mounts small and allows optimization for both areas separately.

When using shared file systems, such as NFS, it is easy to ensure that the standby host has access to the files of all the active hosts in the system. However, proper fencing is not provided throughout all versions of NFS. Starting with version 4, a lease-time based locking mechanism is available, which can be used for I/O fencing. Older versions do not support locking as required for high availability; therefore, other techniques are necessary. Usually the STONITH ("shoot the other node in the head") method is implemented to achieve proper fencing capabilities. Nevertheless, even in NFS version 4 environments, STONITH is commonly used as this may speed up failover on the one hand and it ensures that locks are *always* freed on the other hand.

In such a setup, the Storage Connector API can be used for invoking the STONITH calls. During failover, the SAP HANA master host calls the STONITH method of the custom Storage Connector with the hostname of the failed host as input value. Usually, hardware partners maintain a mapping of hostnames to management network addresses, which is used to send a power cycle signal to the server via management network or cluster manager. When the host comes up again, it will automatically start in standby host role. This behavior can be illustrated as follows:



Scenario:

The worker host 2 fails, e.g. is not responding to pings of the master node. It is not known whether the host has died completely or there is just a network failure. It cannot be determined if the host still is able to write to the storage.

SAP HANA's behavior with an active custom Storage Connector:

1. The master host pings the worker host and does repeatedly not receive an answer within a certain timeout.
2. The master host decides that the standby host 3 shall take over host 2's role and initiates the failover.
3. The master host calls the custom Storage Connector with the hostname of the failing host as parameter. The custom Storage Connector sends a power cycle request to its management entity, which in turn triggers a power cycle command to the failing host 2.
4. Only after the custom Storage Connector returns without error, the standby is entitled to acquire the persistences of the failed host and proceeds with the failover process.

SAP hardware partners and their storage partners are responsible for developing a corruption-safe failover solution.

A common alternative shared file system is IBM's GPFS, which handles fencing properly: the local disks of each SAP HANA server are used to setup clustered file system that is shared amongst all servers. By putting a certain level of redundancy to this cluster, the file system can cope with disk failures.

Disaster Recovery Approaches

Backups

Backups are a low-cost approach to disaster recovery with less extensive storage requirements than other approaches like storage or system replication.

There are two types of backups. *Data backups* can be triggered manually or scheduled. A data backup effectively copies a database savepoint that is stored as a snapshot to the backup destination location. A *log backup* occurs automatically when a log segment (a log segment is represented by a file on disk with a fixed size) fills up or a configurable time threshold is exceeded. The log segment is copied to the backup destination. This may happen in parallel to a data backup.

Usually, data and log backups are backed up to the same storage device.

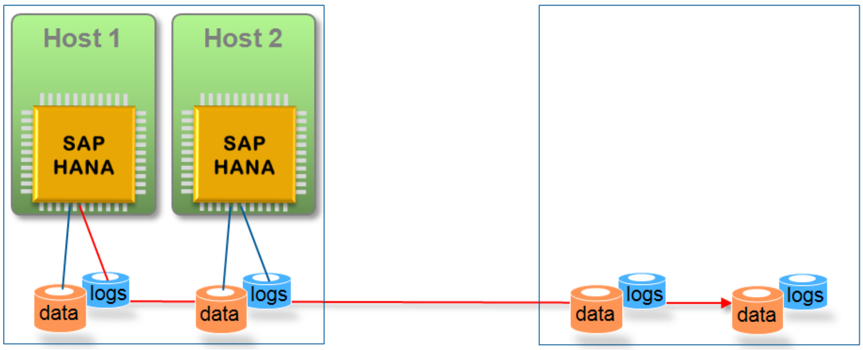
SAP HANA offers the following options to configure the backup storage:

- *External backup tools via the BACKINT interface*: data and log backups are transferred to a third-party provider software, which transports the data and log backups to another location.

- *Shared file system*: an external shared file system is mounted to all hosts of a distributed landscape ensuring that even after host auto-failovers the correct locations are accessible to the hosts.

Storage Replication

Storage replication setup is transparent to SAP HANA: the storage subsystem mirrors (replicates) all write operations to another remote storage device, either synchronously or asynchronously depending on the device type, distance, and customer preference. Connecting multiple sites (in a chain or even star layout) is also possible and used quite often.



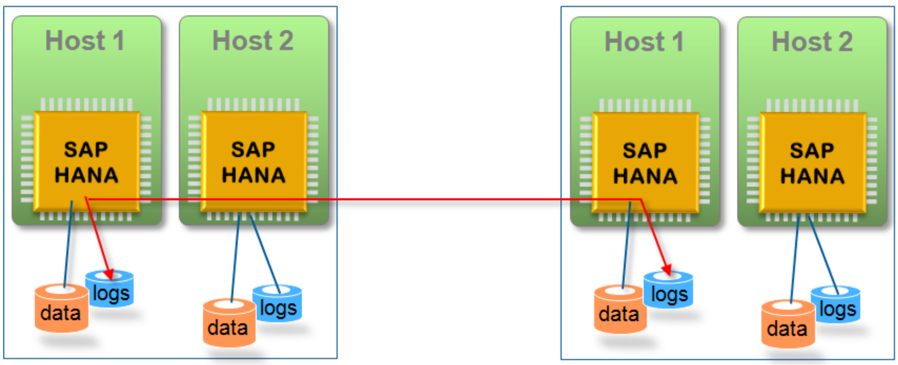
In synchronous mode, a redo log entry will only be considered complete when it is persisted on both sites. This results in a higher latency than without replication.

Another important condition of storage replication is that the correct write *ordering* must be ensured across the entire SAP HANA system. This condition may impact the storage solution.

Upon failure of the primary site, a failover is made to the secondary site, which uses the replicated storage. Note that administrative care must be taken to carefully fence off the failed system in case of a takeover. Usually, an external cluster manager is employed to ensure this kind of fencing.

System Replication

In this approach, each SAP HANA instance communicates with a corresponding peer in the secondary system to persist the same data and logs as on the primary system.



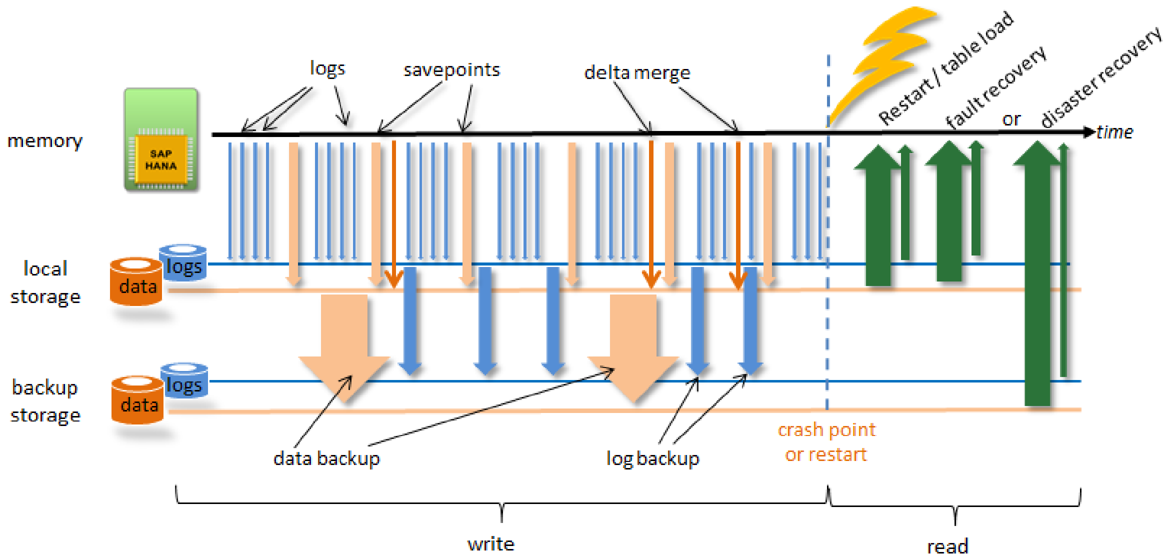
This solution is hardware and vendor agnostic. The two systems are symmetric (with symmetric storage devices), to allow failover and failback operations to occur symmetrically. Write ordering is not a concern due to SAP HANA's internal roll-forward mechanisms during failover. The replication will be either synchronous or asynchronous depending mainly on the distance between the two sides but does not impose any particular requirements on the storage solution.

Note that here as well administrative care must be taken to fence off a failed system, which usually is achieved by employing an external cluster manager.

Furthermore, so-called *near zero downtime upgrades* of SAP HANA are supported with this approach by updating the secondary site first, issuing a takeover to this site, re-register the old primary as new secondary system and start the whole procedure again for the old primary.

Performance

The I/O performance of the storage system impacts specific scenarios which occur during the life cycle of a SAP HANA database. The performance of the transactional load, backup & recovery, delta merge or a recovery after a disaster depends on the data throughput and latency of the storage system. In the following picture some basic I/O activity between memory and storage is illustrated and will be explained in greater detail in the below section on *scenarios*.



Scenarios

The scenarios that are influenced by the I/O performance of the storage subsystem are described in greater detail in the following table.

Scenario	Description
Write Transactions	All changes to data are captured in the redo log. HANA asynchronously persists the redo log with I/O orders of 4 KB to 1 MB size into log segment files in the log volume. Transactions writing a commit into the redo log wait until the buffer containing the commit has been written to the log volume.
Savepoint	A savepoint ensures that all changed persistent data since the last savepoint is written to the data volume. The SAP HANA database triggers savepoints in 5 minutes intervals by default on each service individually. Data is automatically persisted from memory to the data volume located on disk. Depending on the type of data the block sizes vary between 4 KB and 16 MB. Savepoints run asynchronously to HANA update operations. Database update transactions only wait at the critical phase of the savepoint, which is taking a few microseconds.
Snapshot	The SAP HANA database snapshots are used by certain operations like backup and database copy. They are created by triggering a system wide transactional consistent savepoint. The system keeps the persistent data belonging to the snapshot at least until the drop of the snapshot. In the *SAP HANA Administration Guide* [7] more detailed information can be found about snapshots.

Scenario	Description
Delta Merge	The delta merge takes place in memory. Updates on column store tables are stored in the in-memory delta storage. During the delta merge these changes are compressed and applied to the in-memory main storage optimized for reads. Right after the delta merge, the new main storage is persisted to the data volume. The delta merge does not block parallel read and update transactions. Please refer to [7] for details on the delta merge.
Database restart	At database startup the services load their row store tables and the database catalog from the storage subsystem into memory. Additionally, the redo log entries persisted after the last savepoint are read from the log volume and replayed in memory.
Column store table load	Column store tables are loaded on demand into memory. The last loaded columns before the stop of the system are known and asynchronously reloaded after restart.
Failover (Host Auto-Failover)	On the standby host the services are running in idle mode. Upon failover, the data and log volumes of the failed host are automatically assigned to the standby host, which then has read and write access to the files of the failed active host. Row as well as column store tables (on demand) are loaded into memory. The log entries are replayed.
Takeover (Storage Replication)	The takeover of the secondary site in storage replication works just like the database start (see above).
Takeover (System Replication)	The secondary system is already running, i.e. the services are active but do not accept SQL and thus are not usable by applications. Just like the database restart (see above) the row store tables get loaded into memory from the storage subsystem. If table preload is used, then most of the column store tables are already in memory as they were loaded in the primary site. During takeover the replicated redo logs that were shipped since the last data transfer from primary to secondary must be replayed.
Online Data Backup	For a data backup the current payload of the data volumes is read and copied to the backup location. For further information on backups please refer to [7].
Database Recovery	During the restore of the database, first the data backup is read from the backup location and written into the SAP HANA data volumes. The I/O write orders of this data recovery have a size of 64 MB. Also, the redo log can be replayed during a database recovery, i.e. the log backups are read from the backup location and the log volumes and the log entries get replayed.
Queries	No access to the storage subsystem is required. Queries might force column store table loads, if the respective tables are not loaded into memory yet.

I/O Patterns

In the following table the I/O order sizes and patterns are listed resulting from the mentioned scenarios. Operations which have a similar I/O pattern are summarized in one table row.

Scenarios	Data Volume	Redo Log Volume	Backup Storage
Write transactions		**WRITE** *OLTP* – mostly 4 KB sequential I/O orders; *OLAP* – larger I/O order sizes (up to 1 MB)	

Scenarios	Data Volume	Redo Log Volume	Backup Storage
Savepoint, Snapshot, Delta merge	**WRITE** 4 KB – 64 MB asynchronous parallel I/O orders (amount of data depends on system load)		
DB Restart, Failover, Takeover	**READ** 4 KB – 64 MB asynchronous parallel I/O orders (amount of data depends on RowStore size)	**READ** 256 KB asynchronous I/O orders	
Column store table load	**READ** 4 KB – 16 MB asynchronous parallel I/O orders		
Data Backup	**READ** 4 KB – 64 MB asynchronous I/O orders copied into buffers of 512 MB		**WRITE** 512 MB sequential I/O orders (configurable)
Log Backup		**READ** 4 KB – 128 MB asynchronous I/O orders copied into buffers of 128 MB	**WRITE** 4 KB – 128 MB sequential I/O orders
Database Recovery	**WRITE** 4 KB – 64 MB asynchronous parallel I/O orders	**READ** 256 KB asynchronous I/O orders	**READ** *Data backup*: 512 MB I/O orders copied into buffers of 512 MB *Log backup*: 128 MB I/O orders copied into buffers of 128 MB
Queries			

I/O Sizing

Storage systems used for SAP HANA must fulfill a certain set of KPIs for minimum data throughput and maximum latency time. In the course of both the SAP HANA appliance and the SAP HANA Enterprise Storage certification process, the fulfillment of those KPIs is checked using the SAP HANA Hardware and Cloud Measurement Tools (HCMT). All KPIs must be met for each SAP HANA compute node connected to a given storage system.

Storage vendors may use different approaches to reach those KPIs for their storage families. The maximum number of SAP HANA compute nodes connected to a given storage system may differ accordingly. The vendor-specific storage configuration document for SAP HANA systems describes in detail how many SAP HANA nodes can be connected in parallel and how to configure the storage system for optimal collaboration with SAP HANA.

For SAP HANA TDI environments, the KPIs of the Enterprise Storage certification apply which are available as part of the official HCMT documentation [9]. The KPIs applicable for the certification of SAP HANA appliances are communicated to the SAP HANA Hardware Partners only.

For more details about SAP's Enterprise Storage certification please refer to the SAP HANA TDI [8].

Summary

SAP HANA offers a wide range of configuration possibilities for different kind of storage subsystems. One option is to use a SAP HANA appliance, which is based either on shared or non-shared storage approach depending on the choice of the hardware partner providing a fully integrated high availability and performance concept. Another option is to use an existing storage subsystem within a tailored datacenter integration approach. For this, the performance and the high availability capabilities must be certified by SAP.

The sizing of the storage depends on the sizing of the main memory and the number of hosts that belong to the system as well as on the influence of the I/O performance on the mentioned typical SAP HANA database scenarios. Generally spoken, if all hosts of a system have a certain size of main memory, it is recommended that the storage subsystem provides at least three times this amount of persistent space. Depending on the backup strategy, a minimum of half of this amount or more must be added. If a disaster recovery solution (system or storage replication) is added to the overall system, the storage requirements basically double.

The performance requirements that must be fulfilled by the storage subsystem mainly rely on the scenario that is driven by the SAP HANA database. In generally, OLAP scenarios are usually optimized for fast loading times and OLTP scenarios are optimized for low latency log writing.

In terms of design, to accommodate automatic local fail-over to a standby host, a solution for storage sharing with appropriate fencing must be employed. SAP HANA offers several design choices and provides a ready to use solution for the non-shared storage case. If this solution cannot be applied, because either the storage does not fulfill the requirements of the Storage Connector or a shared storage approach is used, SAP HANA offers the Storage Connector API for storage technology vendors and hardware partners to develop a corruption-safe failovering solution.

Finally, SAP HANA offers different approaches to data replication (to cope with disasters, but also with logical errors as well as to have the opportunity to easily apply the productive data to other SAP HANA systems).

Which solutions eventually will be used at customer site highly depends on the requirements of data redundancy and RPO (recovery point objective) as well as RTO (recovery time objective). Please refer to [6] for further details.

Terminology Appendix

Fencing

"Fences out" an entity of a distributed system that is not acting normally. Usually, this entity will be killed or all shared resources will be revoked from it.

Host Auto Failover

The **Master host** coordinates transactions and governs the system topology. There is only one master at a time.

A **Standby host** is a passive component of the system. It has all services running, but not data volumes assigned waiting for failure of others to take over their role.

A **Worker host** is an active component accepting and processing requests.

LUN

Logical Unit Number – an identifier of a storage device

SAP HANA Appliance

A pre-configured black-box HANA system including everything from storage over network to server layer. Several hardware partners offer such systems in different sizes.

SCSI-3 Persistent Reservations

A built-in mechanism of the SCSI-3 protocol, which is widely supported by most storage subsystems. Based on registered keys, a device can be reserved, i.e., locked.

Split Brain

A situation in a distributed system where more than one host demands the master role for itself, usually because the connection is broken between them.

System Replication

The **Primary site** is an active HANA system accepting and processing requests.

A **Secondary site** is a passive copy of the primary system, which has some parts of the data preloaded into memory (SAP HANA ≤ SPS07) but is not accepting any kind of requests.

Tailored datacenter integration

Next to the black-box appliance approach, a customer can choose and/or re-use hardware to save costs, to have more flexibility according to his or hers IT landscape and to optimize for special requirements.

References

[1] Sizing SAP HANA
https://help.sap.com/docs/SAP_HANA_PLATFORM/eb3777d5495d46c5b2fa773206bbfb46/d4a122a7bb57101493e3f5ca08e6b039.html

[2] SAP Quick Sizer
https://www.sap.com/about/benchmark/sizing.sizing-guidelines.html#quick-sizer

[3] SAP Note 1793345: Sizing for SAP Suite on HANA
https://me.sap.com/notes/1793345

[4] SAP Note 1872170: Suite on HANA memory sizing report
https://me.sap.com/notes/1872170

[5] SAP Note 1736976: Sizing Report for BW-on-HANA
https://me.sap.com/notes/1736976

[6] SAP HANA High Availability White Paper
https://www.sap.com/documents/2016/05/f8e5eeba-737c-0010-82c7-eda71af511fa.html

[7] SAP HANA Administration Guide
https://help.sap.com/docs/SAP_HANA_PLATFORM/6b94445c94ae495c83a19646e7c3fd56/330e5550b09d4f0f8b6cceb14a64cd22.html

[8] SAP HANA Tailored Data Center Integration
https://www.sap.com/documents/2017/09/e6519450-d47c-0010-82c7-eda71af511fa.html

[9] SAP Note 2493172: SAP HANA Hardware and Cloud Measurement Tools
https://me.sap.com/notes/2493172

[10] The Delta Merge Operation
https://help.sap.com/docs/SAP_HANA_PLATFORM/6b94445c94ae495c83a19646e7c3fd56/bd9ac728bb57101482b2ebfe243dcd7a.html