Vblock® Systems 740
SAP HANA tailored data center integration
Best practices

Document revision 1.1
September 2015
## Revision history

<table>
<thead>
<tr>
<th>Date</th>
<th>Version</th>
<th>Description of changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>August 2015</td>
<td>1.1</td>
<td>Minor updates based on iDev review and feedback</td>
</tr>
<tr>
<td>May 2015</td>
<td>1.0</td>
<td>Initial version</td>
</tr>
</tbody>
</table>
RAID considerations ..................................................................................................................... 37
Thin pools...................................................................................................................................... 38
Meta volumes for data and log.......................................................................................................38
EMC VMAX masking view................................................................................................................38
   Initiator group .......................................................................................................................... 38
   Port group ............................................................................................................................... 38
   Storage group ......................................................................................................................... 38
Accessing EMC VMAX storage from the SAP HANA nodes........................................................39
   Native Linux multipathing .......................................................................................................39
   XFS file system ....................................................................................................................... 41
   Linux logical volume management ..........................................................................................42
   SAP HANA storage connector API ..........................................................................................42

EMC VMAX3 configuration recommendations .................................................................................43
   Fast elements .......................................................................................................................... 43
   Rotational speed requirements ............................................................................................... 43
   Virtual provisioning data devices ........................................................................................... 43
   Storage resource pool .............................................................................................................. 43
   Service level objective .......................................................................................................... 44
   Host connectivity .................................................................................................................... 45
   FA-director/port requirements ................................................................................................. 45
   EMC VMAX 3 masking view ................................................................................................. 47
      Initiator group ..................................................................................................................... 47
      Port groups ......................................................................................................................... 48
      Storage group ...................................................................................................................... 49
   Sizing guidelines ..................................................................................................................... 49
Introduction

SAP HANA tailored data center integration enables a customer to use existing enterprise compute, network and storage components, which already exist in their datacenter, as opposed to purchasing a dedicated appliance for SAP HANA.

This guide outlines how SAP HANA TDI can be implemented on a Vblock® System 740. SAP HANA TDI - enabled by VCE - is designed for rapid technology expansion and scaling of the Vblock System compute, network, and storage resources. SAP HANA TDI allows VCE to offer HANA directly with its standard Vblock, VxBlock, and VxRack platforms. It provides the benefits of the VCE Experience - built, delivered, and supported by VCE.

For additional information about terminology, refer to the VCE Glossary.

Purpose

This guide contains best practice guidelines for implementing SAP HANA TDI on an existing Vblock System.

Audience

The audience for this document consists of members of the VCE Professional Services organization and certified partner organizations authorized to perform an SAP HANA TDI implementation on behalf of VCE.
What is a TDI-ready Vblock® System?

A TDI-ready Vblock System is a standard Vblock System that uses SAP HANA certified components and can be configured for HANA and made ready for the SAP HANA TDI validation process. Since it is a standard Vblock System, no changes are expected during the physical manufacturing process. Changes to the logical configuration can be provided by VCE Professional Services or during the logical build. Currently, configuration changes are provided by either VCE or partner Professional Services. Future versions might embed some configuration changes during Vblock System physical or logical configuration.

Network interface requirements

The default Vblock System network interface is defined as follows:

- Management Network (1 or 10 GB): eth0 (vNIC0) - Communication with the management network (for example, VLAN 105) for Infrastructure and SAP HANA administration.
  
  **Note:** This is defined between application servers, storage and AMP and is not used for anything other than Vblock System infrastructure management.

- HANA Internode Network (1 or 10 GB): eth1 (vNIC1) - Communication between all HANA nodes and NFS access (e.g., VLAN 109)

- Customer Network (1 or 10 GB): eth2 (vNIC2) - Communication with the customer/production network VLAN (customer specific)

Networks for SAP HANA

Different types of network communication channels can be defined to support the different SAP HANA scenarios and setups depending on customer needs. These include:

- Client zone networks
- Internal zone networks
- Storage zone networks
- Infrastructure-related networks
**Client zone networks**

Channels used for external access to SAP HANA functions by end-user clients, administration clients, and application servers, and for data provisioning through SQL.

<table>
<thead>
<tr>
<th>Name</th>
<th>Use case</th>
<th>Solutions</th>
<th>Required bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application server network</td>
<td>Communication between SAP application server and database</td>
<td>All</td>
<td>1 or 10 GB Ethernet</td>
</tr>
<tr>
<td>Client network</td>
<td>Communication between SAP application server and database</td>
<td>All</td>
<td>1 or 10 GB Ethernet</td>
</tr>
<tr>
<td>Data source network</td>
<td>Communication between SAP application server and database</td>
<td>Optional for all SAP HANA systems</td>
<td>1 or 10 GB Ethernet</td>
</tr>
</tbody>
</table>

**Internal zone networks**

Channels used for SAP HANA internal communication within the database or, in a distributed scenario, for communication between hosts.

<table>
<thead>
<tr>
<th>Name</th>
<th>Use case</th>
<th>Solutions</th>
<th>Required bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internode network</td>
<td>Node-to-node communication within a scale-out configuration</td>
<td>Scale-out</td>
<td>10 GB Ethernet</td>
</tr>
<tr>
<td>System replication network</td>
<td>Communication between SAP application server and database</td>
<td>SAP HANA Disaster Tolerance(DT)</td>
<td>To be defined with customer</td>
</tr>
</tbody>
</table>

**Storage zone networks**

Channels used for storage access (data persistence) and for backup and restore procedures.

<table>
<thead>
<tr>
<th>Name</th>
<th>Use case</th>
<th>Solutions</th>
<th>Required bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backup network</td>
<td>Data backup</td>
<td>Optional for all SAP HANA systems</td>
<td>10 GB Ethernet or 8 GB/s Fibre Channel network</td>
</tr>
<tr>
<td>Storage network</td>
<td>Communication between SAP application server and database</td>
<td>SAP HANA Disaster Tolerance(DT)</td>
<td>10 GB Ethernet or 8 GB/s Fibre Channel network</td>
</tr>
</tbody>
</table>
### Infrastructure related networks

<table>
<thead>
<tr>
<th>Name</th>
<th>Use case</th>
<th>Solutions</th>
<th>Required bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Management network</td>
<td>Infrastructure and SAP HANA administration</td>
<td>Optional for all SAP HANA systems</td>
<td>1 GB Ethernet</td>
</tr>
</tbody>
</table>

### NIC requirement per blade

The Cisco UCS B460 M4 Blade Server can fit four adaptors. However, adapter slot 1 is dedicated for only the Cisco UCS Virtual Interface Card (VIC) 1240. No other mezzanine card fits into adapter slot 1.

To double the bandwidth throughput (an increase of 40 GB/s), add a port expander to the Cisco UCS VIC 1240. Add an additional 80 GB/s by mixing a Cisco UCS VIC 1240 and VIC 1280 in the same server.

For maximum performance, use a Cisco UCS VIC 1240 + port expander + VIC 1280 to allow for both adapter redundancy and 160 GB/s of bandwidth to the blades. Note that to get to 160 GB/s you need all
16 uplinks between the two Cisco UCS 2208XP Fabric Extenders and fabric interconnects, as shown in the following diagram:

![Diagram of Cisco UCS 5108 Blade Server Chassis](image)

**Figure 1: Configuration for maximum bandwidth**

Use the Cisco UCS VIC 1240, port expander, and Cisco UCS VIC 1280 to provide high amounts of bandwidth to blades. This provides the flexibility between price and performance. The VIC cards internally connect inside the Cisco UCS 5108 Server Chassis to the Cisco UCS 2100/2200 Series Fabric Extender IOM modules, which are then uplinked to the fabric interconnects (for example Cisco UCS 6248UP Fabric Interconnect). Depending on your bandwidth and redundancy needs, there are several options to choose from:

<table>
<thead>
<tr>
<th>Bandwidth</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>40 GB/s</td>
<td>Cisco UCS VIC 1240</td>
</tr>
<tr>
<td>80 GB/s</td>
<td>Cisco UCS VIC 1280 or Cisco UCS VIC 1240 + Port Expander</td>
</tr>
<tr>
<td>160 GB/s</td>
<td>Cisco UCS VIC 1280 and Cisco UCS VIC 1240 + Port Expander</td>
</tr>
</tbody>
</table>
Disk sizing requirements

The number of disks required for the HANA persistence depends on the disk type (10K or 15K RPM or enterprise flash drive (EFD), capacity requirements and RAID protection. A mirrored (RAID-1) protection in the storage array provides the best performance for applications with heavy write activities such as SAP HANA. This applies primarily to 10K RPM and 15K RPM drives. EFDs can be configured as RAID-5, either 3+1 or 7+1 (3+1 might offer higher availability, especially for large EFDs).

To meet the host IOPS (input/output per second) requirements on 10K or 15K RPM disks, distribute the HANA persistence across a certain number of disks. A HANA worker host generates approximately 1,200 IOPS. A 10K RPM HDD supports approximately 120 IOPS and a 15K RPM HDD supports approximately 150 IOPS.

For example, in the 6+1 HANA scale-out installation (six worker hosts and one standby host, 7,200 total host IOPS) and one-tier storage configuration, distribute the persistence across at least 60 x 10K RPM or 48 x 15K RPM disks for HANA persistence. Additional disk capacity might be required for OS, HANA shared file system (performance is not important for this) is required. Choose the disk size that meets the capacity requirements and provides the best total cost of ownership (TCO).
## Solution architecture for SAP HANA® scale-out design

This section describes the solution design for an appliance-like SAP HANA scale-out implementation using the following components:

| Compute          | • Intel Ivy Bridge EX E7 4890v2  
|                  | • Cisco UCS B460 M4 Blade Server  
|                  | • Cisco UCS 6248UP Fabric Interconnect  
|                  | • Cisco UCS 6296UP Fabric Interconnect  
|                  | • Cisco UCS 2204XP Fabric Extender IOM module  
|                  | • Cisco UCS 2208XP Fabric Extender  
| Network          | • Cisco Nexus 3064-T Switch  
|                  | • Cisco Nexus 9396PX Switch or Cisco Nexus 5548UP Switch or Cisco Nexus 5596UP Switch  
|                  | • Cisco MDS 9148S Multilayer Fabric Switch or Cisco MDS 9706 Multilayer Director Switch  
| Storage          | • EMC VMAX - 10K, 20K, 40K  
|                  | • EMC VMAX3 - 100K, 200K, 400K  
| Operating systems| • SUSE Linux Enterprise Server  
|                  | • Red Hat Enterprise Linux 6.5 or higher  
| SAP HANA         | • SPS 09, Rev 90 or greater  

An SAP HANA TDI solution on Vblock Systems offers multiple HANA worker nodes and additional standby nodes. A SAP HANA worker node is a database server belonging to a SAP HANA scale-out system with its own data and log file systems. These file systems are stored on the storage system/array. A SAP HANA standby node is used for high availability (HA). If a worker node fails, its HANA data and log file systems fail over to the standby node and the standby becomes a worker node. The default number of standby nodes per production HANA system is one.
The following diagram shows the SAP HANA TDI solution on the Vblock System 740:

![Vblock System 740 SAP HANA diagram](image)

*Figure 2: SAP HANA TDI Vblock System 740 design*
Network configuration

Cisco UCS defaults meet SAP's core requirements for SAP HANA. Cisco UCS is based on 10 GB Ethernet and provides redundancy through the dual-fabric concept. The following diagram shows Cisco UCS network paths:

Each Cisco UCS chassis is linked through four 10 GB Ethernet connections to each Cisco UCS fabric interconnect. Those southbound connections can be configured in port channel mode or pinning mode, the pinning mode was used for better control of the network traffic. The service profile configuration helps ensure that through normal operation, all traffic in the internal zone is on fabric A, and all other traffic (client zone and storage zone) is on fabric B. Management traffic is also on fabric A. This configuration helps ensure that the network traffic is distributed across both fabrics.

Internode traffic flows only from the blade to the fabric interconnect and back to the blade. All other traffic must travel over the Cisco Nexus 5500 platform switches to the storage resource or to the data center network.

The storage array uses fibre channel for the data and log volumes, two host bus adaptors (HBA) must also be configured: one per fabric. The multipath driver for SUSE Linux Enterprise Server is used for path redundancy and to distribute traffic over both fabrics as required. With the integrated algorithms for
bandwidth allocation and quality of service (QoS), Cisco UCS and Cisco Nexus switches help provide the best traffic distribution.

**Maintenance policies**

VCE recommends defining a maintenance policy with **Reboot Policy** set to **User Ack** for the SAP HANA server.

This policy helps ensure that a configuration change in Cisco UCS does not automatically force all SAP HANA servers to restart. The administrator must acknowledge the restart for the servers changed in Cisco UCS; otherwise, the configuration change takes effect when the server restarts through an OS command.

**Configuring the SAN boot**

For SAN boot, configure virtual host bus adapters (vHBA) configured. The storage WWPN is required. As a best practice, the SAN zoning should allow each HBA to see two controllers, storage processors, and front-end ports of the boot storage.

Change the boot policy to enable SAN boot. If the boot LUN should be reachable over multiple paths, the SAN boot configuration should list the WWPNs of the storage mapped to the two vHBAs on the server.

Any change in boot policy requires a server restart to write the new setting in the BIOS. Without a SAN boot device configuration in the BIOS, the VIC Fibre Channel boot driver is not loaded and SAN boot is not possible. If the SAN LUN is successfully detected, the KVM console displays the following message: **Option ROM installed successfully.**

The system is ready to start the OS installation. The configuration in the OS installation procedure must be adopted, following the single-path or multipath design of the boot LUN.

**SAP HANA storage access**

This section presents basic information about the configuration of SAP HANA storage at the OS level. The underlying infrastructure configuration is discussed earlier in this document.

The information presented here is a high-level overview on best practices for the configuration process.

**Related information**

[Solution architecture for SAP HANA® scale-out design](#) (see page 11)

**Block storage for SAP HANA data and log files**

The block storage configuration in the OS for data and log files is the same for all Fibre Channel solutions regardless of whether the traffic traverses native Fibre Channel or FCoE.
Use a multipath configuration for block storage. This configuration using the VCE solution for SAP HANA and native Linux multipathing (device manager multipath I/O [DM-MPIO]) on the SAP HANA nodes improves performance and provides high availability for the access paths to the storage devices.

**File storage for /hana/shared**

The SAP HANA data and log volumes are based on a shared-nothing model. In addition, SAP HANA requires a shared area in which all SAP HANA nodes for one security identifier (SID) have access all time in parallel: /hana/shared/<SID>.

**Note:** The size of the /hana/shared file system must be at least equal to the main memory of all SAP HANA nodes, and the file system type must be able to expand the size if a new node is added to the SID.

**Shared network**

This section describes the VCE solution for the SAP HANA TDI implementation option for a shared network.

Using SAP HANA TDI, organizations can run multiple SAP HANA production systems in one VCE solution, creating an appliance-like solution. Many customers already use this option for their non-production systems. Another option is to run the SAP application server using the SAP HANA database on the same infrastructure as the SAP HANA database.

Another option is to install a SAP HANA database in an existing Cisco UCS deployment used to run SAP and non-SAP applications.
Multiple SAP HANA SIDs in one HANA TDI setup

With SAP HANA TDI, you can run multiple SAP HANA systems in the same infrastructure solution, as shown in the following diagram:

![Diagram of two SAP HANA SIDs in one SAP HANA TDI setup](image)

**Figure 4: Two SAP HANA SIDs in one SAP HANA TDI setup**

**Requirements**

Using multiple SAP HANA SIDs in one appliance requires one file system for /hana/shared per SID. For fibre channel-based solutions, change the LUN mapping so that only the servers for a specific SID can see the data and log LUNs.

**Additional options**

Additional options include dedicated VLAN IDs per SID and QoS settings per VLAN.
SAP HANA and SAP Application Server in one HANA TDI setup

You can run the SAP HANA-related SAP application server on the same infrastructure as the SAP HANA database. With this configuration, the solution controls the communication between the application server and the database. This approach quarantines the bandwidth and latency for best performance and includes the application server in the disaster tolerance solution together with the SAP HANA database.

**Figure 5: SAP HANA TDI: database and application in one HANA TDI setup**

**Requirements**

Using SAP HANA and the SAP application server in one appliance requires a dedicated server for the SAP applications. Use the same server type as for the SAP HANA database (Cisco UCS B460 M4), or add servers such as the Cisco UCS B200 M4 Blade Server to run the application directly on the blade or as a virtualized system with a supported hypervisor.
The storage for the OS and application can be hosted on the same external storage as used for SAP HANA. However, this setup can degrade SAP HANA database performance on this storage, so separate storage is recommended.

**Additional options**

Additional options include dedicated VLAN IDs and QoS settings per VLAN. You can introduce a dedicated application to the database network based on VLAN separation.
Storage best practice guidelines

SAP HANA is an in-memory database. The data is kept in the RAM of one or multiple SAP HANA worker hosts and all database activities such as reads, inserts, updates, or deletes are performed in the main memory of the host and not on disk. This differentiates SAP HANA from other traditional databases, where only a part of the data is cached in RAM and the remaining data resides on disk.

Note: In this document, EMC VMAX storage is referred to as follows:

- VMAX, also known as the VMAX Family, consists of 10K, 20K, and 40K storage arrays.
- VMAX3, also known as the VMAX3 Family consists of 100K, 200K, and 400K storage arrays.

This document addresses the storage best practice guidelines and recommendations for the VMAX and VMAX3 Families. It indicates where a procedure or best practice is relevant for only one family.

Scale-up versus scale-out

<table>
<thead>
<tr>
<th>Environment</th>
<th>Database</th>
<th>Benefit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale-up</td>
<td>Needs to fit into RAM of a single server</td>
<td>Preferred environment for online transaction processing workloads, such as SAP Business Suite</td>
</tr>
</tbody>
</table>
| Scale-out   | Tables are distributed across RAM of multiple servers | Use worker and standby hosts  
• A worker host is an active component and accepts and processes database requests.  
• A standby host is a passive component with all database services running, but no data in RAM. It is waiting for a failure of a worker host to take over its role. This process is called host auto-failover. |

As in-memory capacity can be very high, scale-out HANA clusters are perfectly suited for online analytical processing (OLAP) type workloads with very large data sets.

SAP HANA TDI scalability

SAP HANA TDI scalability defines the number of production HANA worker hosts (in scale-out installations) or single hosts (in scale-up installations) that can be connected to enterprise storage arrays and still meet the SAP performance key performance indicators (KPI) for enterprise storage. Because the disk capacity for the HANA persistence is always related to the HANA database RAM capacity, the
required disk capacity for multiple HANA hosts is not the limiting factor in most cases. Enterprise storage arrays can provide much more capacity than required for HANA. The scalability depends on several other factors. For example:

- Array model, cache size, disk types
- Bandwidth, throughput, and latency
- Overall use and resource consumption of the array
- How the HANA host is connected to the array
- Storage configuration of the HANA persistence

**Note:** The scalability numbers in this document for the EMC VMAX and VMAX3 arrays are recommendations based on performance tests on various models. The actual number of HANA hosts that can connect to a VMAX and VMAX3 in a customer environment can be higher or lower than the number of HANA hosts referred to in this document. Use the SAP HANA hardware configuration check tool (HWCCT) tool to validate SAP HANA performance and determine the maximum possible number of HANA hosts on a given storage array.

**SAP HANA persistence**

SAP HANA uses disk storage for the following purposes:

- To maintain the persistence of in-memory data on disk to prevent data loss due to a power outage and to allow a host auto-failover, where a standby HANA host takes over the in-memory data of a failed worker host in scale-out installations
- To log information about data changes (redo log)

Each SAP HANA worker host (scale-out) or single-host (scale-up) requires two file systems on disk storage: a data file system and a log file system.

**Capacity considerations**

The required capacity for the SAP HANA persistence on disk depends on the in-memory database size and the HANA server RAM size.

SAP HANA users must perform memory and CPU sizing as the first step to sizing a SAP HANA deployment.

- New SAP HANA implementations should size memory and CPU using the HANA version of the SAP Quick Sizer tool, available on the SAP Service Marketplace website or consult SAP for assistance.
- For systems that are migrating to SAP HANA, SAP provides tools and reports for proper HANA memory sizing.
After determining memory requirements, estimate disk capacity requirements by using the sizing rules in the SAP white paper, [SAP HANA Storage Requirements](#). File systems must include the following:

<table>
<thead>
<tr>
<th>Component</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATA</td>
<td>$1.2 \times \text{RAM}$</td>
</tr>
<tr>
<td>LOG less than or equal to 512 GB</td>
<td>$\frac{1}{2} \times \text{RAM}$</td>
</tr>
<tr>
<td>LOG greater than 512 GB</td>
<td>$\times \text{RAM}$</td>
</tr>
<tr>
<td>SHARED</td>
<td>$\times \text{RAM}$ per every 4 worker nodes</td>
</tr>
<tr>
<td>BOOT</td>
<td>100 GB</td>
</tr>
</tbody>
</table>

SAP refers to RAM size as the size of the database instead of as the physical memory size of the servers. For example, the HANA database can consume 1.3 TB RAM on a single host but the host has 2 TB physical RAM capacity. SAP recommends the sizing based of the actual database size, in this example 1.3 TB.

Shared storage can be allocated on any storage arrays that are accessible to all the HANA nodes. On EMC VMAX3, it is allocated through the eNAS (embedded NAS) feature.

**Note:** SAP sizing requirements do not consider future growth of the database.

To calculate the required usable storage capacity for the HANA persistence of a scale-up (single-host) or scale-out (multi-host) appliance, the following details are required:

- RAM size of a HANA worker host
- Number of HANA worker hosts

For example, use the following formulas to calculate the required capacity for a 5+1 HANA scale-out appliance where each server has 1 TB RAM:

<table>
<thead>
<tr>
<th>Total capacity for</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATA</td>
<td>$1.2 \times (S) \times (N) = 1.2 \times 1 \text{ TB} \times 5 = 6 \text{ TB}$</td>
</tr>
<tr>
<td>LOG</td>
<td>$512 \text{ GB} \times (N) = 512 \text{ GB} \times 5 = 2.5 \text{ TB}$</td>
</tr>
<tr>
<td>HANA persistence</td>
<td>$6 \text{ TB} + 2.5 \text{ TB} = 8.5 \text{ TB}$</td>
</tr>
<tr>
<td>SHARED</td>
<td>$(S) \times (N/4) = 1 \text{ TB} \times 2 = 2 \text{ TB}$</td>
</tr>
<tr>
<td>BOOT</td>
<td>$(100 \text{ GB}) \times (N+1) = 100 \text{ GB} \times 6 = 600 \text{ GB}$</td>
</tr>
<tr>
<td>HANA</td>
<td>$\approx 11 \text{ TB}$</td>
</tr>
</tbody>
</table>

Use any free capacity or add additional capacity to accommodate file systems with no performance requirements, such as operating system LUNs.

EMC VMAX storage allows use of a dedicated storage disk group to separate the HANA workload from a non-HANA workload on a shared storage array. Use a dedicated disk group if the impact of non-HANA applications on shared disks results in HANA hosts not being able to meet the performance requirements.
Related information

SAP HANA shared file system on VMAX (see page 23)
SAP HANA OS images on VMAX (see page 23)

Disk type considerations

Due to the specific workload of the SAP HANA database with primarily write I/Os, use the following disk types:

- 7.2k rpm disks
- 10k rpm disks
- 15k rpm disks
- EFD Enterprise flash disks (EFD)

VMAX

For SAP HANA on VMAX 10K, 20K, or 40K arrays, use a single tier (drive type/technology and RAID protection) strategy. This is because all writes to VMAX storage are sent to VMAX persistent cache and are later written to the disk media. The HANA write workload benefits primarily from VMAX cache prior to any fully automated storage tiering (FAST) and multi-tier advantages. A single tier strategy provides an adequate solution for HANA and simplifies deployment.

VMAX 10K, 20K or 40K storage allows you to separate the HANA workload from a non-HANA workload on a shared storage array by using a dedicated storage disk group. A dedicated disk group can be used if the impact of non-HANA applications on shared disks is too high so that the HANA hosts will no longer meet the performance requirements. This is not a requirement and the HANA devices can also reside on shared disks.

VMAX3

With VMAX3 100K, 200K, or 400K storage, service level objective provisioning and host limits control performance for certain applications. FAST technology enhancements allow combining of 7.2k, 10k or 15k rpm hard disk drives (HDDs) and EFDs. The number of disks required depends on the disk type (7.2k, 10k, or 15k rpm or EFD), capacity requirements, and RAID protection. A mirrored (RAID-1) protection in a VMAX array provides the best performance for applications with heavy write activities, such as SAP HANA. This applies primarily to 10k rpm and 15k rpm drives. Configure EFDs as RAID-5, either 3+1 or 7+1 (3+1 offers higher availability, especially for large EFDs). Choose the disk size that meets capacity requirements and provides the best total cost of ownership (TCO).

IOPS

To meet host IOPS requirements on 7.2k, 10k, or 15k rpm disks, distribute the HANA persistence across a certain number of disks. A HANA worker host generates approximately 1,200 I/O operations per second.
IOPS. A 10k rpm HDD supports approximately 120 IOPS and a 15k rpm HDD supports approximately 150 IOPS.

For example, in the 6+1 HANA scale-out installation (six worker hosts and one standby host, 7,200 total host IOPS) and one-tier storage configuration, distribute the persistence across at least 60 x 10k rpm or 48 x 15k rpm disks. 15 TB of usable capacity is required for the HANA persistence. The following table compares the usable capacity of the different disk sizes and the RAID (mirrored) protection.

The following table compares the usable capacity of the different disk sizes and the RAID (mirrored) protection.

<table>
<thead>
<tr>
<th>Disk size</th>
<th>Usable capacity per disk</th>
<th>Usable capacity with 60 disks</th>
<th>Usable capacity mirrored</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>300 GB</td>
<td>268 GB</td>
<td>16,080 GB</td>
<td>8,40 GB</td>
<td>Does not meet capacity requirements</td>
</tr>
<tr>
<td>400 GB</td>
<td>366 GB</td>
<td>21,960 GB</td>
<td>10,980 GB</td>
<td>Does not meet capacity requirements</td>
</tr>
<tr>
<td>600 GB</td>
<td>536 GB</td>
<td>32,160 GB</td>
<td>16,080 GB</td>
<td>Meets capacity requirements with the best TCO. Future growth is limited.</td>
</tr>
<tr>
<td>900 GB</td>
<td>820 GB</td>
<td>49,200 GB</td>
<td>24,600 GB</td>
<td>Meets capacity requirements and enables future growth.</td>
</tr>
</tbody>
</table>

**SAP HANA OS images on VMAX**

You can boot SAP HANA nodes from either local disks or from SAN and VMAX3 devices. If you boot from a SAN, follow the best practices documented in the “Booting from SAN” section of the *EMC Host Connectivity Guide for Linux*.

The capacity required for the operating system is approximately 100 GB per HANA host (worker and standby) and includes capacity for the /usr/sap directory.

**SAP HANA shared file system on VMAX**

In an SAP HANA scale-out implementation, install the SAP HANA database binaries on a shared file system that is exposed to all hosts of a system under a /hana/shared mount point. If a host needs to write a memory dump (which can read up to 90% of RAM size), it will be stored in this file system. Guided by the specific customer infrastructure and requirements, the options for the file systems are:

- VMAX block storage can create a shared file system using a cluster file system such as an Oracle Cluster File System 2 (OCFS2) on top of the block LUNs. SUSE Linux provides OCFS2 capabilities with the high availability package, and a SUSE license is required. The high availability package is also part of the SUSE Linux Enterprise Server (SLES) for SAP applications distribution from SAP that is used by most of the HANA appliance vendors.

- NAS systems, such as EMC VNX, can be used instead of OCFS2 to provide an NFS share for the HANA shared file system.

- eNAS of the VMAX3 arrays can provide the NFS share.
The size of the HANA shared file system should be the total RAM size of the database, which is the number of worker hosts multiplied with the RAM size of a single node. *SAP HANA Storage Requirements* provides more details.

**OCFS2 with SUSE Linux Enterprise Server 11**

Oracle Cluster File System 2 (OCFS2) is a general-purpose journal file system that has been fully integrated since the SUSE Linux 2.6 Kernel. OCFS2 allows storing application binary files, data files, and databases on devices on shared storage. All nodes in a cluster have concurrent read and write access to the file system.

**Setting up a cluster for OCFS2**

**About this task**

This procedure describes how to set up a cluster with existing hosts and add a OCFS2 resource.

**Before you begin**

Set up a cluster for OCFS with SLES 11, as it is not supported in kernel mode. Check for high availability. Verify that firewall is disabled. Do the passwordless login thing (ssh-keygen and ssh-copy-id) for all nodes to all nodes.

**Procedure**

1. Copy `/etc/corosync/corosync.conf.example` to `/etc/corosync/corosync.conf` and edit it to suit your needs.

   **Note:** Be sure to use the network address instead of the IP address in the network. The network address typically ends in .0. (for example, 192.168.187.0.)

2. Start OpenAIS by typing the following command:

   ```
   /etc/init.d/openais start
   ```

3. Use the password `hacluster`. Set a password by typing:

   ```
   passwd hacluster
   ```

4. Type `hb_gui`. The click **Connect**.
5 Verify by typing

```shell
crm_mon -r
```

Both hosts appear in green, meaning the cluster is running.

![Cluster setup screenshot](image)

*Figure 6: Example of cluster setup*

What to do next

Configure OCFS2 services and STONITH resource.
Troubleshooting

- Ensure that high availability packages are installed. The packages needed for configuring and managing a cluster are included in the high availability installation pattern, available with the SUSE Linux Enterprise High Availability Extension.

- Confirm that the high availability extension is installed as an add-on to SUSE Linux Enterprise Server 11 SP3 on each cluster node and that the high availability pattern is installed on each of the machines.

- Ensure that the initial configuration is the same for all cluster nodes. When communicating with each other, all nodes belonging to the same cluster need to use the same `bindnetaddr`, `mcastaddr`, and `mcastport`.

- Verify that all communication channels and options configured in `/etc/corosync/corosync.conf` are the same for all cluster nodes.

- If using encrypted communication, verify that the `/etc/corosync/authkey` file is available on all cluster nodes.

- All `corosync.conf` settings with the exception of `nodeid` must be the same. `authkey` files on all nodes must be identical.

- Ensure the firewall allows communication via the `mcastport`, which is used for communication between the cluster nodes. If it is blocked by the firewall, the nodes cannot see each other.

- To ensure the `mcastport` is not blocked by the firewall, check the settings in `/etc/sysconfig/SUSEfirewall2` on each node. Alternatively, start the YaST firewall module on each cluster node.

Configuring OCFS2 services and a STONITH resource

Before creating OCFS2 volumes, configure the following resources as services in the cluster: DLM, O2CB and a STONITH resource. OCFS2 uses the cluster membership services from Pacemaker, which run in user space. Therefore, DLM and O2CB need to be configured as clone resources that are present on each node in the cluster.

The following procedure uses the `crm` shell to configure the cluster resources. Alternatively, use the Pacemaker GUI to configure the resources. Both cLVM and OCFS2 need a DLM resource that runs on all nodes in the cluster and therefore usually is configured as a clone. In a setup that includes both OCFS2 and cLVM, configuring one DLM resource for both OCFS2 and cLVM is enough.

Configuring DLM and O2CB resources

About this task

This procedure describes how to configure DLM and O2CB resources. The configuration consists of a base group that includes several primitives and a base clone. Both base group and base clone can be used in various scenarios afterwards (for both OCFS2 and cLVM, for example).
Before you begin

Note that you only need to extend the base group with the respective primitives as needed. As the base group has internal co-location and ordering, this facilitates overall setup as there is not need to specify several individual groups, clones, and their dependencies. Complete the following steps for one node in the cluster.

Procedure

1. Start a shell and log in as root or equivalent.
2. Type `crm configure`.
   
   **Note:** The dlm clone resource controls the distributed lock manager service and makes sure this service is started on all nodes in the cluster. Due to the base group’s internal co-location and ordering, the o2cb service is only started on nodes where a copy of the dlm service is already running.
3. Create the primitive resources for DLM and O2CB by typing:
   
   ```bash
   primitive dlm ocf:pacemaker:controld
   op monitor interval="60" timeout="60"
   primitive o2cb ocf:ocfs2:o2cb
   op monitor interval="60" timeout="60"
   ```
4. Create a base group and a base clone by typing:
   
   ```bash
   group base-group dlm o2cb
   clone base-clone base-group
   meta interleave="true"
   ```

   Review your changes by typing:
   
   ```bash
   show
   ```
5. If everything is correct, submit your changes by typing:
   
   ```bash
   commit
   ```

   To exit the crm live configuration, type
   
   ```bash
   exit
   ```

What to do next

Create an SBD (STONITH Block Devices) partition.
Creating an SBD partition

About this task

This procedure describes how to configure a STONITH resource. Without a STONITH mechanism (like external/sbd) in place, the configuration will fail. It is recommended to create a 1 MB partition at the start of the device. If the SBD device resides on a multipath group, adjust the timeouts SBD uses, as MPIO's path down detection can cause latency.

Before you begin

Make sure the SBD device does not hold any data as SBD command overwrites the device without requesting confirmation.

Remember: The SBD partition in the following steps is referred to as /dev/SBD. Replace it with the actual pathname, for example: /dev/sdc1.

Procedure

1. Initialize the SBD device by typing:

```bash
sbd -d /dev/SBD create
```

This writes a header to the device, and creates slots for up to 255 nodes sharing this device with default timings. To use more than one device for SBD, provide the devices by specifying the -d option multiple times. For example:

```bash
sbd -d /dev/SBD1 -d /dev/SBD2 -d /dev/SBD3 create
```

2. If the SBD device resides on a multipath group, adjust the timeouts SBD uses. Use the -4 option is to specify the msgwait timeout. Use the -1 option is to specify the watchdog timeout. These can be specified when the SBD device is initialized (all timeouts are given in seconds). In the following example msgwait is set to 180 seconds and watchdog is set to 90 seconds.

```bash
/usr/sbin/sbd -d /dev/SBD -4 180 -1 90 create
```

3. Type the following command to check what was written to the device:

```bash
sbd -d /dev/SBD dump
```

Header version : 2
Number of slots : 255
Sector size : 512
Timeout (watchdog) : 5
Timeout (allocate) : 2
Timeout (loop) : 1
Timeout (msgwait) : 10

4. Type `crm configure`. 

---

VCE CONFIDENTIAL
Subject to Confidentiality Use Restrictions
5 Configure external/sdb as fencing device with /dev/sdb2 being a dedicated partition on the shared storage for heartbeating and fencing. To do this, type:

```
primitive sbd_stonith stonith:external/sbd 
meta target-role="Started" 
op monitor interval="15" timeout="15" start-delay="15"
```

6 Review your changes by typing:

```
show
```

7 If everything is correct, submit your changes by typing:

```
commit
```

8 Exit the crm live configuration by typing:

```
exit
```

Troubleshooting

- Start (or enable) operation includes checking the status of the device. If the device is not ready, the STONITH resource will fail to start.

- At the same time the STONITH plugin is asked to produce a host list. If this list is empty, there is no point in running a STONITH resource that cannot shoot anything. The host name on which STONITH is running is filtered from the list, since the node cannot shoot itself.

- To use single-host management devices such as lights-out devices, ensure the STONITH resource is not allowed to run on the node that it is supposed to fence. Use an infinitely negative location node preference (constraint). The cluster informs you and moves the STONITH resource to another starting place.

- Each STONITH resource must provide a host list. This list can be added to the STONITH resource configuration or retrieved from the device itself from outlet names. stonithd uses the list to find out which STONITH resource can fence the target node. The STONITH resource can shoot (fence) the node only if the node appears in the list.

- If stonithd does not find the node in any of the host lists provided by running STONITH other nodes. If the target node does not show up in the host lists of other stonithd instances, the fencing request ends in a timeout at the originating node.

Related information

[Creating an SBD partition](#) (see page 28)
Creating and formatting OCFS2 volumes

About this task

This procedure describes how to create and format the OCFS2 volume in order to use the resource for the cluster. Execute the following steps on only one cluster node.

Before you begin

Before you begin, prepare the block devices you plan to use for your OCFS2 volumes. Leave the devices as free space. Then create and format the OCFS2 volume with the `mkfs.ocfs2`. The most important parameters for the command are listed in the following table:

<table>
<thead>
<tr>
<th>OCFS2 parameter</th>
<th>Description and recommendation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Volume Label (-L)</td>
<td>A descriptive name for the volume to make it uniquely identifiable when it is mounted on different nodes. Use <code>tunefs.ocfs2</code> to modify the label as needed.</td>
</tr>
<tr>
<td>Cluster Size (-C)</td>
<td>Cluster size is the smallest unit of space allocated to a file to hold the data. For the available options and recommendations, refer to the <code>mkfs.ocfs2</code> main page.</td>
</tr>
<tr>
<td>Number of Node Slots (-N)</td>
<td>The maximum number of nodes that can concurrently mount a volume. OCFS2 creates separate system files, such as the journals, for each node. Nodes that access the volume can be a combination of little-endian architectures (such as x86, x86-64, and ia64) and big-endian architectures (such as ppc64 and s390x). Node-specific files are referred to as local files. A node slot number is appended to the local file. For example: <code>journal:0000</code> belongs to whatever node is assigned to slot number 0. Set each volume's maximum number of node slots when you create it, according to how many nodes expected to concurrently mount the volume. Use <code>tunefs.ocfs2</code> to increase the number of node slots as needed. Note that the value cannot be decreased.</td>
</tr>
<tr>
<td>Block Size (-b)</td>
<td>The smallest unit of space addressable by the file system. Specify the block size when creating the volume. For the available options and recommendations, refer to the <code>mkfs.ocfs2</code> main page.</td>
</tr>
<tr>
<td>Specific Features On/Off (--fs-features)</td>
<td>If you provide a comma separated list of feature flags, <code>mkfs.ocfs2</code> tries to create the file system with those features set according to the list. To turn a feature on, include it in the list. For an overview of all available flags, refer to the <code>mkfs.ocfs2</code> main page.</td>
</tr>
<tr>
<td>Pre-Defined Features (--fs-feature-level)</td>
<td>Allows you to choose from a set of pre-determined file system features. For the available options, refer to the <code>mkfs.ocfs2</code> main page.</td>
</tr>
</tbody>
</table>
• unwritten
• metaecc
• indexed-dirs
• xattr

Procedure

1. Open a terminal window and log in as root.
2. Check if the cluster is online by typing `crm_mon`.
3. Create and format the volume using `mkfs.ocfs2`.
   
   **Note:** For information about the syntax for this command, refer to the `mkfs.ocfs2` main page.
4. To create a new OCFS2 file system on `/dev/sdb1` that supports up to 32 cluster nodes, type

   ```
   mkfs.ocfs2 -N 32 /dev/sdb1
   ```

   For information about the syntax for this command, refer to the `mkfs.ocfs2` main page.

**Mounting OCFS2 volumes**

You can mount an OCFS2 volume manually or using the cluster manager.

**Manually mounting an OCFS2 volume**

**About this task**

This procedure describes the steps for manually mounting an OCFS2 volume.

**Procedure**

1. Open a terminal window and log in as root.
2. Check if the cluster is online by typing `command crm_mon`.
3. Mount the volume from the command line by typing:

   ```
   mount
   ```

   **Note:** If you mount the OCFS2 file system manually for testing purposes, make sure to unmount it again before starting to use it by means of OpenAIS.
Mounting an OCFS2 volume with the Cluster Manager

About this task

This procedure describes how to mount an OCFS2 volume with the Cluster Manager. The procedure uses the crm shell to configure the cluster resources. Alternatively, you can use the Pacemaker GUI to configure the resources.

Procedure

1. Start a shell and log in as root or equivalent.
2. Type `crm configure`.
3. Configure Pacemaker to mount the OCFS2 file system on every node in the cluster:
   ```
   primitive ocfs2-1 ocf:heartbeat:Filesystem \
   params device="/dev/sdb1" directory="/mnt/shared" fstype="ocfs2" options="acl" \
   op monitor interval="20" timeout="40"
   ```
4. Add the file system primitive to the base group you have configured in Configuring DLM and O2CB Resources.
   a. Type:
      ```
      edit base-group
      ```
   b. In the vi (visual editor) that opens, modify the group as follows and save your changes:
      ```
      group base-group dlm o2cb ocfs2-1
      ```
   
   **Note:** Due to the base group’s internal co-location and ordering, Pacemaker starts the `ocfs2-1` resource on nodes that also have an `o2cb` resource already running.
5. Review your changes by typing:
   ```
   show
   ```
6 Verify that you have configured all needed resources.

See the following example:

```bash
primitive clvm ocf:lvm2:clvmd \
    params daemon_timeout="30"
primitive dlm ocf:pacemaker:controld \
    op monitor interval="60" timeout="60"
primitive o2cb ocf:ocfs2:o2cb \
    op monitor interval="60" timeout="60"
primitive ocfs2-1 ocf:heartbeat:Filesystem \
    params device="/dev/sdb1" directory="/mnt/shared" fstype="ocfs2" options="acl" \
    op monitor interval="20" timeout="40"
primitive sbd_stonith stonith:external/sbd \
    meta target-role="Started" op monitor interval="15" \
    timeout="15" start-delay="15"
primitive vgl ocf:heartbeat:LVM \
    params volgrpname="cluster-vg" \
    op monitor interval="60" timeout="60"
group base-group dlm o2cb clvm vgl ocfs2-1
clone base-clone base-group \
    meta interleave="true"
```

7 If everything is correct, submit your changes by typing:

```bash
commit
```

8 Leave the crm live configuration by typing:

```bash
exit
```
Results
The aim is to have a Pacemaker cluster of N nodes and an OCFS2 resource available.

Figure 7: Pacemaker GUI - example

eNAS for HANA share with VMAX3

eNAS (Embedded NAS) extends the value of VMAX3 to file storage by enabling you to leverage vital enterprise features including service level objective based provisioning, host I/O limits, and FAST technologies for both block and file storage. EMC VMAX3 Unified is a multi-controller, transaction NAS solution designed for users requiring hyper consolidation for block storage (the traditional VMAX use case) combined with moderate file storage in mission-critical environments.

For more information on eNAS and VMAX3, refer to:

- VMAX3 Series
- eNAS Technical Overview
EMC VMAX configuration recommendations

The following configuration recommendations apply to SAP HANA production systems deployed on EMC VMAX 10K, 20K, and 40K storage arrays. Production systems in SAP HANA TDI environments must meet SAP performance requirements (KPIs) and the configuration requirements described in this section.

Host connectivity

Front-end director port requirements

Special attention is required when connecting SAP HANA nodes to the front-end director ports (FA ports) of an EMC VMAX array.

On a VMAX director, two FA ports share one dedicated CPU core. For example, FA-1E:0 and FA-1E:1 share the same core. To achieve full I/O performance for production HANA deployments, consider the following FA port requirements for a VMAX array:

- Dedicate FA ports to HANA and do not share them with non-HANA applications.
- Use only one FA port per CPU core on the I/O module and do not use the adjacent port. For example, use FA-1E:0 and leave FA-1E:1 unused. Do not use the adjacent port for non-HANA applications.
- Never connect a single HBA (Host Bus Adapter) to both ports of the same director.
- The minimum number of FA ports required for HANA depends on the number of HANA nodes connected to a single VMAX engine. For example, if 16 HANA nodes are connected to a dual-engine VMAX, use 5 ports on each engine for only HANA. Use the following table to determine the required number of FA ports:

<table>
<thead>
<tr>
<th>HANA worker nodes</th>
<th>Required FA-ports</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - 2</td>
<td>2</td>
</tr>
<tr>
<td>3 - 4</td>
<td>3</td>
</tr>
<tr>
<td>5 - 6</td>
<td>4</td>
</tr>
<tr>
<td>7 - 8</td>
<td>5</td>
</tr>
<tr>
<td>9 - 10</td>
<td>6</td>
</tr>
<tr>
<td>11 - 12</td>
<td>8</td>
</tr>
</tbody>
</table>

- Connect all HANA hosts to all FA ports dedicated to HANA. The more available I/O paths for a host, the better the SAP HANA performance.
- Balance FA-ports used for HANA across all available EMC VMAX engines.
• Use 8 Gb/s FC ports. While 10 Gb/s iSCSI or Fibre Channel over Ethernet (FCoE) can be used, VCE has not validated it for SAP HANA. HANA 2 Gb/s or 4 Gb/s FC ports are not supported.

The following diagrams show the rear view of the VMAX engines with 4-port FC I/O modules (8 Gb/s) for host connectivity. VCE recommends using the I/O ports marked with a yellow box for HANA connectivity. Leave the adjacent ports unused.

![Figure 8: Rear view of a VMAX 10K engine](image)

![Figure 9: Rear view of a VMAX 20K and 40K engine](image)

**EMC VMAX sizing guidelines**

In a 10K, 20K, or 40K EMC VMAX array, the scalability of SAP HANA primarily depends on the number of available engines in the array. Table 4 shows the VMAX models and the estimated maximum number of HANA worker hosts that can be connected according to the number of available engines.

<table>
<thead>
<tr>
<th>VMAX model</th>
<th>Number of available engines</th>
<th>Maximum number of HANA worker hosts</th>
</tr>
</thead>
<tbody>
<tr>
<td>10K</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>30</td>
</tr>
<tr>
<td>20K</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>20</td>
</tr>
</tbody>
</table>
### Virtual provisioning considerations

VMAX arrays use EMC Virtual Provisioning to provide capacity to an application. Capacity is allocated using virtual provisioning data devices and provided in thin pools based on the disk technology and RAID type. Thin devices are host-accessible devices bound to thin pools and natively striped across the pool to provide the highest performance.

**Important:** Virtual provisioning (RAID, thin pools, and meta volumes for data and log) best practice guidelines are for the VMAX Family (10K, 20K, and 40K) only.

#### RAID considerations

To provide best write performance for the HANA persistence, RAID-1 mirrored configurations are required for the TDATs on 10k or 15k rpm disks. You can configure virtual provisioning data devices on EFDs using RAID-5, either 3+1 or 7+1. We recommend 3+1.

---

**Table 2: VMAX 10K, 20K, and 40K scalability**

<table>
<thead>
<tr>
<th>VMAX model</th>
<th>Number of available engines</th>
<th>Maximum number of HANA worker hosts</th>
</tr>
</thead>
<tbody>
<tr>
<td>10K</td>
<td>3</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>68</td>
</tr>
<tr>
<td>40K</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>62</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>82</td>
</tr>
</tbody>
</table>

When using EMC Symmetrix Remote Data Facility (SRDF) for SAP HANA storage replication, a reduced number of front-end FA-ports are available. Adjust the number of HANA worker hosts that can connect to the array accordingly.
Thin pools

VCE recommends creating one thin pool for all SAP HANA data volumes and another thin pool for the HANA log volumes in a VMAX array. However, if a limited number of disks are available in smaller HANA environments, performance could be improved by using a single thin pool for both devices. Thin pools consist of virtual provisioning data devices. The number and size of the virtual provisioning data devices in a thin pool depends on the SAP HANA capacity requirements and must be configured using VMAX configuration best practices.

Meta volumes for data and log

Each HANA worker host requires one data and one log volume for the persistent file systems. The sizes of these volumes depend on the Capacity considerations described earlier in this document.

Related information

Capacity considerations (see page 20)

EMC VMAX masking view

EMC VMAX uses masking views to assign storage to a host. Create a single masking view for each HANA host. A masking view consists of the following components:

- Initiator group
- Port group
- Storage group

Initiator group

Port group

The port group contains the front-end director ports to which the HANA host is connected. For the EMC VMAX 10K, 20K and 40K arrays, the more ports assigned to the HANA hosts, the better the performance. However, ensure that a single HBA connects to only one port per director.

Storage group

An SAP HANA scale-out cluster uses the shared-nothing concept for database persistence, where each HANA worker host uses its own pair of data and log volumes and has exclusive access to the volumes during normal operations. If a HANA worker host fails, the HANA persistence of the failed host is used on
a standby host. This requires that all persistent devices are visible to all HANA hosts because every host can become a worker or a standby host.

The VMAX storage groups of a HANA database must contain all persistent devices of the database cluster. The HANA name server, in combination with the SAP HANA storage connector API (application program interface), takes care of proper mounting and I/O fencing of the persistence.

**Accessing EMC VMAX storage from the SAP HANA nodes**

The SAP HANA database requires a SUSE Linux Enterprise Server 11 or a Red Hat Enterprise Linux 6.5 operating system on the HANA nodes. To access the EMC VMAX block devices from the HANA nodes, ensure that zoning is based on SAN best practices. A single HBA must connect to only one port per director.

**Native Linux multipathing**

To access the block devices from the HANA nodes, first enable native Linux multipathing. Follow the steps described in *EMC Host Connectivity Guide for Linux* to enable Linux DM-multipathing on Red Hat Enterprise Linux 6.5 or SUSE SLES 11.

The following sections provide examples of the content in multipath.conf files.
SUSE Linux Enterprise Server 11 example

```bash
## This is a template multipath-tools configuration file
## Uncomment the lines relevant to your environment
##
defaults {
    # udev_dir /dev
    # polling_interval 10
    # selector "round-robin 0"
    # path_grouping_policy multibus
    # getuid_callout "/lib/udev/scsi_id -g-u-d /dev/%n"
    # prio const
    # path_checker directio
    # rr_min_io 100
    # max_fds 8192
    # rr_weight priorities
    # fallback immediate
    # no_path_retry fail
    user_friendly_names no
}

blacklist {
    ## Replace the wwid with the output of the command MPIO
    ## 'scsi_id -g-u-s /block/[internal scsi disk name]'  
    ## Enumerate the wwid for all internal scsi disks.
    ## Optionally, the wwid of VCM database may also be listed here
    
    wwid 35005076718d4224d
    devnode"^(ram|raw|loop|fd|md|dm-|sr|scd|st)[0-9]*"
    devnode "^hd[a-z][0-9]*"
    devnode "^cciss!c[0-9]*d[0-9]*[p0-9]*"
}
```

Red Hat Enterprise Linux example

```bash
## This is a template multipath-tools configuration file
## Uncomment the lines relevant to your environment
##
defaults {
    # udev_dir /dev
    # polling_interval 10
    # selector "round-robin 0"
    # path_grouping_policy multibus
    # getuid_callout "/sbin/scsi_id -g-u-s /block/%n"
    # prio_callout /bin/true
    # path_checker readsector0
    # rr_min_io 100
    # rr_weight priorities
    # fallback immediate
    # no_path_retry fail
    user_friendly_names no
}

## The wwid line in the following blacklist section is shown as an example
## of how to blacklist devices by wwid. The 3 devnode lines are the
## compiled in default blacklist. If you want to blacklist entire types
## of devices, such as all scsi devices, you should use a devnode line.
## However, if you want to blacklist specific devices, you should use
## a wwid line. Since there is no guarantee that a specific device will
```
## not change names on reboot (from /dev/sdato /dev/sdb for example)

## devnode lines are not recommended for blacklisting specific devices. ##

Remove # to enable the devnode blacklist. You can add the WWID for the Symmetrix VCM database, as shown in this example. The VCM database is a read-only device that is used by the array. Blacklisting it eliminates any error messages that might occur because of its presence.

### Blacklist example

```bash
wwid 3600604800000190101965533030303230
devnode "^\(ram|raw|loop|fd|md|dm-|sr|scd|st\)[0-9]*"
devnode "^hd[a-z]"
devnode "^cciss!c[0-9]d[0-9-]"
```

The HANA persistent devices should be visible on a HANA worker host after a restart or a rescan (command - `rescan-scsi-bus.sh`). Verify that all devices are present and each device has the configured number of active paths by typing the following command: `multipath-ll`

The following is an example of the output generated by typing the `multipath-ll` command.

```
3600009700000298700460533030303238 dm-10 EMC,SYMMETRIX
size=512G features='0' hwhandler='0' wp=rw
  `-+-policy='round-robin 0' prio=1 status=active
     |2:0:5:1 sda7 66:32 active ready running
     |1:0:5:1 sdb7 68:192 active ready running
     `-2:0:4:1 sdcg 69:64 active ready running
3600009700000298700460533030303338 dm-11 EMC,SYMMETRIX
size=512G features='0' hwhandler='0' wp=rw
  `-+-policy='round-robin 0' prio=1 status=active
     |2:0:5:3 sda8 66:64 active ready running
     |1:0:5:3 sdcb 68:224 active ready running
     `-2:0:4:3 sdcj 69:96 active ready running
3600009700000298700460533030303438 dm-12 EMC,SYMMETRIX
size=1.5T features='0' hwhandler='0' wp=rw
  `-+-policy='round-robin 0' prio=1 status=active
     |2:0:5:5 sdam 66:96 active ready running
     |1:0:5:5 sdcc 69:0 active ready running
     `-2:0:4:5 sdcl 69:128 active ready running
3600009700000298700460533030303538 dm-14 EMC,SYMMETRIX
size=1.5T features='0' hwhandler='0' wp=rw
  `-+-policy='round-robin 0' prio=1 status=active
     |2:0:5:6 sdan 66:112 active ready running
     |1:0:5:6 sdcd 69:16 active ready running
     `-2:0:4:6 sdcl 69:144 active ready running
```

### XFS file system

The XFS file system provides the best performance for both HANA data and log block devices.
To format a block device with the XFS file system, type the following command on the HANA node:

```
mkfs.xfs /dev/mapper/3600009700002987004605330303238
```

Run this command for all block devices. If for some reason a file system must be expanded, use the `xfs_growfs` command on the Linux host after the volume has been expanded on the EMC VMAX array.

**Linux logical volume management**

You can use logical volume management (LVM) on the HANA host to manage devices in a more flexible way. This document assumes that all HANA persistent devices are presented to the HANA hosts as a single device and that LVM is not required.

Logical volume management can address challenges faced in environments needing more flexibility where the size of HANA persistent devices has to be adjusted in more granular increments than available with a MetaLUN expansion on the VMAX. LVM requires the use of a special storage connector API (fcClientLVM), which is part of the SAP HANA software distribution.

**SAP HANA storage connector API**

In an SAP HANA scale-out environment with worker and standby nodes, the SAP HANA storage connector API for Fibre Channel (fcClient) mounts and unmounts the devices to the HANA nodes. If LVM is used, it requires a special version of the API (fcClientLVM).

In addition to mounting the devices, the storage connector API also writes SCSI-3 PR (persistent reservations) to the devices using the Linux command: `sg_persist`

This is called I/O fencing and ensures that at a given time only one HANA worker host has access to a set of data and log devices.
EMC VMAX3 configuration recommendations

The following configuration recommendations apply to production SAP HANA production systems deployed on EMC VMAX3 100K, 200K, and 400K enterprise storage arrays. Production systems in SAP HANA TDI environments must meet SAP performance requirements (KPIs) and the configuration requirements outlined in the following sections.

Fast elements

EMC Fully Automated Storage Tiering (FAST) automates the identification of active or inactive application data for reallocating that data across different pools within a VMAX3 storage array. FAST proactively monitors workloads to identify busy data that would benefit from being moved to higher-performing drives, while also identifying less-busy data that could be moved to higher-capacity drives, without affecting existing performance. This promotion/demotion activity is based on achieving service level objectives that set performance targets for associated applications, with FAST determining the most appropriate pool to allocate data on.

With VMAX3, the following storage elements are pre-configured for ease of manageability and cannot be changed:

- Storage disk groups
- Virtual provisioning data devices
- Data pools
- Storage resource pool

Rotational speed requirements

SAP HANA requires 15k or 10k rpm HDDs or EFDs. HDDs with 7.2k rpm drives do not meet SAP HANA performance requirements.

Virtual provisioning data devices

Each disk group is pre-configured with virtual provisioning data devices based on EMC best practices for size and RAID protection. SAP HANA requires RAID-1 (mirrored) on HDDs and RAID-5 3+1 or 7+1 on EFDs. For EFDs, RAID-5 3+1 is best.

Storage resource pool

A storage resource pool is a collection of data pools that make up a FAST domain. A data pool can only be included in one storage resource pool. EMC VMAX3 ships with a single pre-configured storage resource pool. EMC support is required for custom configurations.
The following diagram shows a sample EMC VMAX3 configuration and a single storage resource pool with multiple disk groups and data pools. In larger HANA environments and where the separation of the HANA workload is required, VCE recommends using a dedicated storage resource pool for the HANA devices. If HANA is installed on a multi-tier storage resource pool, use service level objective (SLO) provisioning to ensure that HANA data is allocated on a higher tier.

![Figure 10: VMAX3 FAST elements](image)

**Service level objective**

In EMC VMAX3 arrays, FAST technology delivers service level objective performance levels. Thin devices can be added to storage groups and storage groups can be assigned to specific service level objectives to set performance expectations. The service level objective defines the response time for the storage group. FAST continuously monitors and adapts the workload to maintain (or meet) the response time target.

There are five available service level objectives, varying in expected average response time targets. An additional optimized service level objective has no explicit response time target associated with it. The following table lists the available service level objectives:

<table>
<thead>
<tr>
<th>Service level objective</th>
<th>Behavior</th>
<th>Expected average response time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diamond</td>
<td>Emulates EFD performance</td>
<td>0.8 ms</td>
</tr>
<tr>
<td>Service level objective</td>
<td>Behavior</td>
<td>Expected average response time</td>
</tr>
<tr>
<td>-------------------------</td>
<td>--------------------------------------------------------------------------</td>
<td>--------------------------------</td>
</tr>
<tr>
<td>Platinum</td>
<td>Emulates performance between EFD and 15 k rpm drives</td>
<td>3.0 ms</td>
</tr>
<tr>
<td>Gold</td>
<td>Emulates 15 k rpm performance</td>
<td>5.0 ms</td>
</tr>
<tr>
<td>Silver</td>
<td>Emulates 10 k rpm performance</td>
<td>8.0 ms</td>
</tr>
<tr>
<td>Bronze</td>
<td>Emulates 7.2 k rpm performance</td>
<td>14.0 ms</td>
</tr>
<tr>
<td>Optimized (default)</td>
<td>Achieves optimal performance by placing most active data on higher performance storage and least active data on most cost-effective storage</td>
<td>N/A</td>
</tr>
</tbody>
</table>

The actual response time of an application associated with each service level objective varies based on the application workload. It depends on average I/O size, read/write ratio, and the use of local or remote replication.

If the HANA devices are created on a dedicated storage resource pool with just EFDs and/or 10k or 15k rpm HDDs, then select optimized service level objective. If not using a dedicated storage resource pool, select at least a platinum service level objective to ensure that data is allocated on EFDs, and/or select 10k or 15k rpm disks.

To further refine response time expectations, add one of the four workload types shown in the following table to the selected SLO (except for optimized).

<table>
<thead>
<tr>
<th>Workload</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLTP</td>
<td>Small block I/O workload</td>
</tr>
<tr>
<td>OLTP with replication</td>
<td>Small block I/O workload with local or remote replication</td>
</tr>
<tr>
<td>Decision support system</td>
<td>Large block I/O workload</td>
</tr>
<tr>
<td>Decision support system with replication</td>
<td>Large block I/O workload with local or remote replication</td>
</tr>
</tbody>
</table>

To improve the latency for small 4K I/O operations on HANA log devices, assign the OLTP workload type to the HANA storage group.

**Host connectivity**

The SAP HANA nodes connect to the EMC VMAX3 arrays through a Fibre Channel SAN. All SAN components require 8 GB/s or 16 GB/s link speed and the SAN topology should follow best practices with all redundant components and links.

**Front-end director port requirements**

Because CPU cores are dynamically allocated to FA director ports in EMC VMAX3 arrays, you can connect HANA hosts to any port on a VMAX3 director. However, connect (or zone) a single HBA initiator...
to only one FA port per director. You can achieve increased availability and performance by connecting each HANA node to different directors, and by using multiple host initiators.

**Note:** You will see no performance or availability benefits from connecting the same host initiator to multiple ports on the same director. If connecting the same host initiator to multiple ports on the same FA, contact VCE support to enable the VMAX3 Fixed Block Architecture (FBA).

To achieve full I/O performance for production HANA deployments, consider the following FA port requirements for the VMAX3 array:

- Dedicate FA-ports to HANA and do not share them with non-HANA applications.
- Do not connect a single HBA port to more than one port on the same director.
- Use the following table to determine the required number of FA-ports, which can be distributed across the available engines. The number of FA ports required for HANA depends on the number of HANA nodes connected to a single VMAX3 engine.

<table>
<thead>
<tr>
<th>HANA worker nodes</th>
<th>Required FA-ports</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 - 4</td>
<td>2</td>
</tr>
<tr>
<td>5 - 8</td>
<td>4</td>
</tr>
<tr>
<td>9 - 16</td>
<td>8</td>
</tr>
<tr>
<td>17 - 20</td>
<td>12</td>
</tr>
</tbody>
</table>

- Distribute FA ports used for HANA across all available VMAX3 engines and balance them between directors. For example, if 16 HANA nodes are connected to a VMAX3 with two engines, balance the connectivity across the engines. Use eight ports on each engine (four per director) for HANA only.
- Use 8 GB/s or 16 GB/s FC ports.
- Ensure that zoning between host initiators and storage ports does not cross switches and uses ISL.
The following diagram shows the rear view of the EMC VMAX3 engine. Each engine has two directors with up to 16 FC front-end ports (ports 4-11 and 24-31).

Figure 11: Rear view of an EMC VMAX3 engine with FA port assignments

Related information

Port groups (see page 48)

**EMC VMAX 3 masking view**

EMC VMAX3 uses masking views to assign storage to hosts. Create a single masking view for each HANA host. A masking view consists of the following components:

- Initiator group
- Port group
- Storage group

**Initiator group**

The initiator group contains the port WWN (PWWN) initiators from the HBAs of the HANA host. Connect each HANA host to the EMC VMAX array with at least two HBAs. Initiator groups can be cascaded and a masking view initiator group can be a collection of the initiator groups from each HANA node.
Port groups

The port group contains all the EMC VMAX3 front-end ports to which the HANA host is connected. If the VMAX3 Enable Dual Port FBA flag is set as described in Front-end director port requirements, then all FA ports used by HANA can be defined in a single port group. However, if this flag is not set, use multiple port groups to ensure that a single HBA port connects to only one FA port per director.

The following diagram shows an environment with 12 HANA hosts. Each host with two HBAs connects with dual fabric to the FA-ports of a dual engine EMC VMAX3.

![Port groups diagram](image)

**Figure 12: EMC VMAX3 SAN connectivity and port groups**

In the VMAX3 SAN connectivity and port groups diagram, we created two port groups. Each port group contains four FA-ports (a total of 8 for 12 HANA hosts), balanced across two engines and the two directors per engine. Port group PG01 is used by HANA hosts hana01-hana06. Port group PG02 is used by hana07-hana12. This connectivity ensures that a single HBA connects to only one port per director.

All 12 HANA hosts belong to the same HANA database scale-out cluster. Therefore, all HANA persistent devices (data and log) belong to the same VMAX storage group. If the HANA hosts belong to multiple clusters, then one storage group per HANA cluster is required. The port group assignment does not change even with multiple HANA clusters.
**Storage group**

An SAP HANA scale-out cluster uses the shared-nothing concept for database persistence, where each HANA worker host uses its own pair of data and log volumes and has exclusive access to the volumes during normal operations. If a HANA worker host fails, the HANA persistence of the failed host is used on a standby host. This concept requires that all persistent devices be visible to all HANA hosts, because every host can become a worker or a standby host.

**Sizing guidelines**

In a 100K, 200K, or 400K VMAX3 array, SAP HANA scalability primarily depends primarily on the number of available engines in the array. The following table shows the EMC VMAX3 models and the estimated maximum number of HANA nodes that can be connected according to the number of available number of engines:

<table>
<thead>
<tr>
<th>EMC VMAX3 model</th>
<th>Engines</th>
<th>Maximum HANA nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>100K</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>200K</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>52</td>
</tr>
<tr>
<td>400K</td>
<td>1</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>32</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>44</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>56</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>104</td>
</tr>
</tbody>
</table>

If using SRDF for SAP HANA storage replication, a reduced number of front-end FA ports are available. Adjust the maximum number of HANA worker hosts that can be connected to the array accordingly.
About VCE

VCE accelerates the adoption of converged infrastructure and cloud-based computing models that dramatically reduce the cost of IT while improving time to market for enterprises and service providers globally. Through its leading Vblock Systems, VCE delivers the industry's only true converged infrastructure, leveraging Cisco compute and network technology, EMC storage and data protection, and VMware virtualization and virtualization management. VCE solutions are available through an extensive partner network and cover horizontal applications, vertical industry offerings and application development environments, enabling customers to focus on business innovation instead of integrating, validating, and managing IT infrastructure.

For more information, go to [http://www.vce.com](http://www.vce.com).