CONCEPTS GUIDE BlueBoxx

Version 2.1

SPHINX | ASPERNBRÜCKENGASSE 2 | 1020 WIEN

TEL: +43 1 599 31 - 0 | FAX: +43 1 599 31 - 99 | info@blueboxx.at

www.blueboxx.at | www.sphinx.at

BlueBoxx Legal Notices

The information contained herein is subject to change without notice and is not warranted to be error-free. If you find any errors, please report them to us in writing.

This software or hardware is developed for general use in a variety of information management applications. It is not developed or intended for use in any inherently dangerous applications, including applications that may create a risk of personal injury.

Sphinx is a registered trademark of Sphinx IT Consulting GmbH and/or its affiliates. Other names may be trademarks of their respective owners.

Intel and Intel Xeon are trademarks or registered trademarks of Intel Corporation. All SPARC trademarks are used under license and are trademarks or registered trademarks of SPARC International, Inc. AMD is a trademark or registered trademark of Advanced Micro Devices. Unix is a registered trademark of The Open Group. Red Hat and Red Hat Enterprise Linux are trademarks or registered trademarks of Red Hat Inc. Microsoft, Microsoft Windows, the Microsoft Windows logo and Microsoft Windows Server are trademarks or registered trademarks of Microsoft Corporation. Oracle, Oracle VM, Oracle VM Server, Oracle VM Manager and Oracle Enterprise Linux are trademarks or registered trademarks of the Oracle Corporation. OpenSuse and Suse Enterprise Server are trademarks or registered trademarks of SUSE LLC.

This software or hardware and documentation may provide access to or information about content, products, and services from third parties. Sphinx IT Consulting GmbH and its affiliates are not responsible for and expressly disclaim all warranties of any kind with respect to third-party content, products, and services unless otherwise set forth in an applicable agreement between you and Sphinx. Sphinx IT Consulting GmbH and its affiliates will not be responsible for any loss, costs, or damages incurred due to your access to or use of third-party content, products, or services, except as set forth in an applicable agreement between you and Sphinx.

Table of Contents

1 PREFACE 4
1.1 Audience 4
1.2 Structure 4
2 OVERVIEW 5
2.1 High Availability 6
2.2 High-Performance 7
2.3 Ready to Use 7
2.4 Cost-effective 7
3 BLUEBOXX ARCHITECTURE 8
3.1 Architecture Overview 8
3.2 Network Architecture 9
3.3 High Availability functionality provided by the BlueBoxx 12
3.4 Data integrity 14
3.5 High Performance 16
3.6 Monitoring 16
3.7 Virtual Shared Storage 18
3.8 Backup 19
3.9 Updates 21
4 HARDWARE 22
4.1 Hardware Requirements 22
4.2 Separation of BlueBoxx nodes 23
5 BENEFITS FOR ORACLE CUSTOMERS 25
5.1 Support and Certification 25
5.2 Licensing Options for Oracle Products 25
6 GLOSSARY 26
7 REFERENCES 26
8 INDEX OF FIGURES 27

1 PREFACE

1.1 Audience

The BlueBoxx Concepts Guide is intended for system administrators and end users who want to learn about the components of the BlueBoxx. This guide outlines the high availability and virtualization principles upon which the BlueBoxx concept is built.

1.2 Structure

The second chapter provides an overview and discusses the strategic aspects of the BlueBoxx. The following chapters contain technical descriptions, explanations of architecture and requirements.

2 OVERVIEW

The BlueBoxx is a high-availability, high-performance and cost effective product with enterprise ready virtualization support out of the box. It is shipped completely pre-configured as a plug-and-play system based on standard x86 hardware and is completely certified by Sphinx IT Consulting GmbH. The BlueBoxx concept hides the storage organization and usage from the compute layer, as shown in Figure 1. The storage and compute layers are connected transparently through the BlueBoxx layer. The BlueBoxx concept enables high availability and high-performance without requiring special purpose hardware.

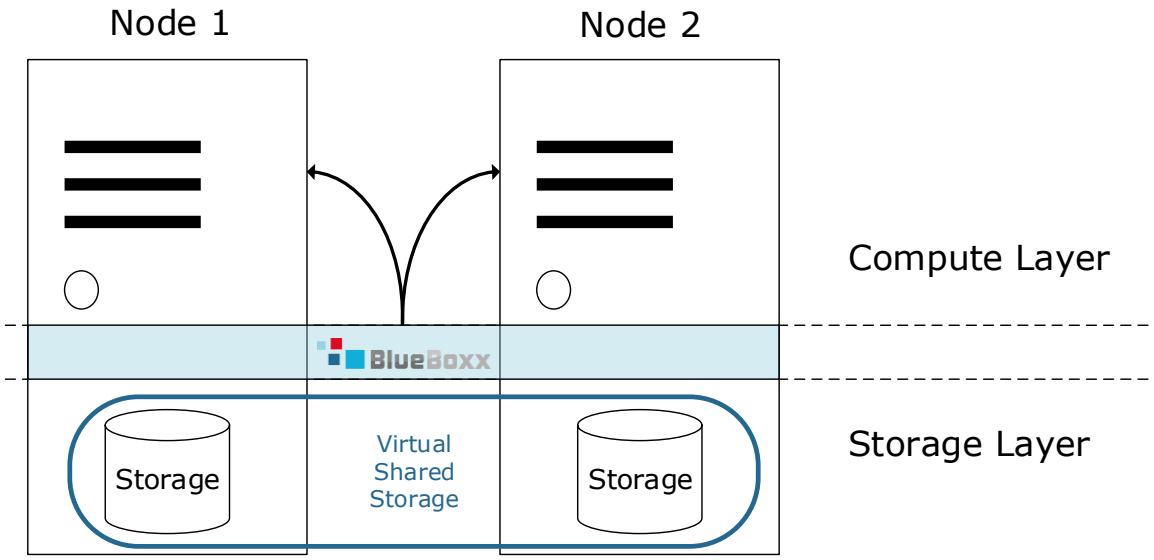

Figure 1 - The BlueBoxx Concept

For the storage layer the customer may choose between internal storage technologies such as:

- Solid-state Drive via serial bus (SATA, SAS, etc.)
- NVMe SSD via PCI Express
- PCI Express Solid-state drives
- Hard Disk Drives (SAS, SATA, etc.)

It is highly recommended to make extensive use of solid-state drive technology. The maximum number of disks is determined by the chosen RAID controllers and the available space for server racks. Furthermore, any external storage connected to the node is considered as local storage, but the latency may be higher when compared to internal storage.

Internal storage must be configured as a RAID. Recommended levels are:

- Mirrored RAID 1
- Block-level striping with distributed parity RAID 5
- Block-level striping with double distributed parity RAID 6

2.1 High Availability

Usually achieving high availability is expensive due to the need for redundant servers, storage, and network infrastructure. The BlueBoxx uses only two identical standard x86 nodes and provides virtual shared storage via the BlueBoxx layer. The BlueBoxx utilizes two layers of hardware redundancy and one layer of application redundancy to achieve high availability across multiple nodes.

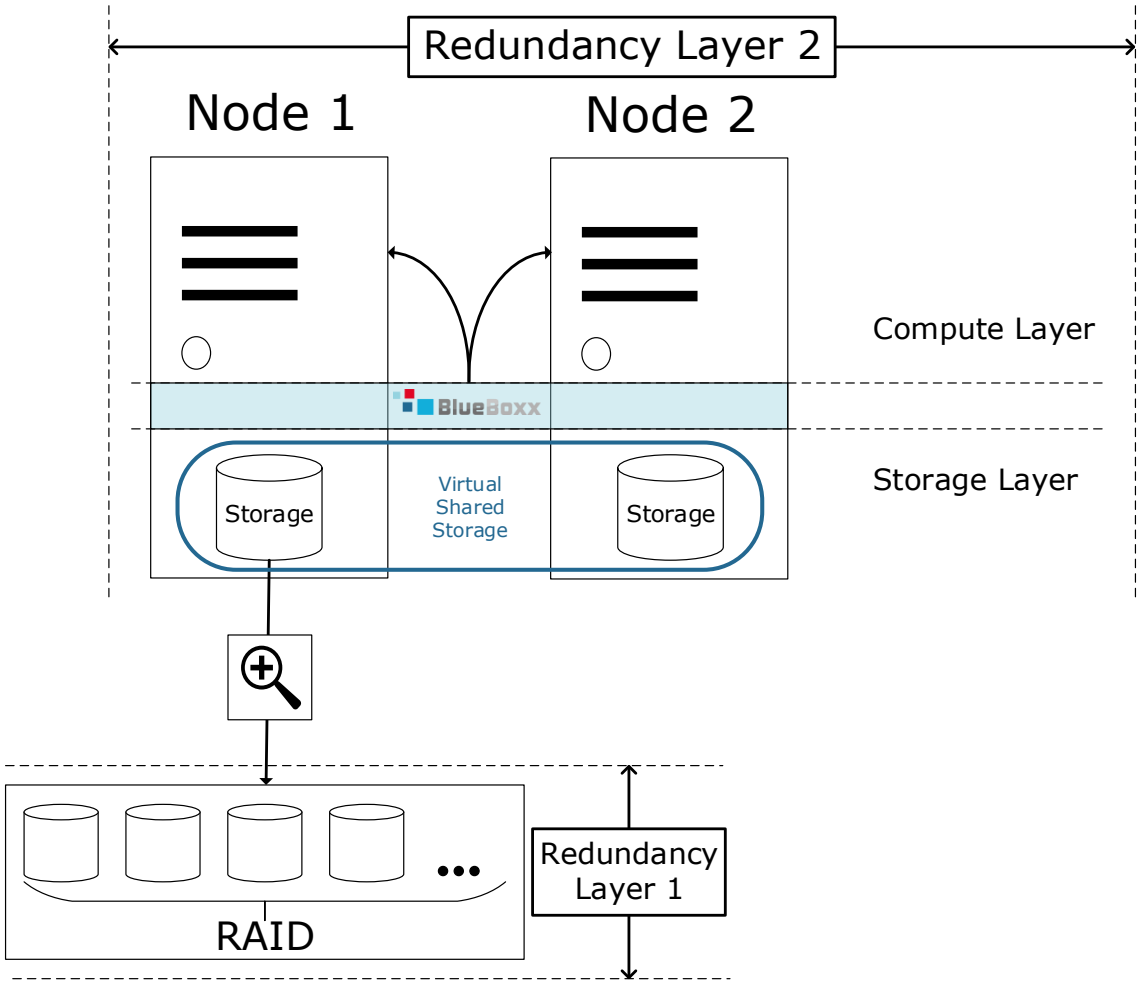

Figure 2 - The two hardware layers of redundancy

The first layer of hardware redundancy relies on independent block devices configured as a RAID array (see chapter 2). Based on the concept of independent nodes, each node has its own hard drives or solid state disks and RAID controllers. The second layer of hardware redundancy spans the two BlueBoxx nodes. If one node is shut down due to maintenance or unexpected outage the second node takes over.

The BlueBoxx layer is responsible for mirroring the data through the connected nodes. Read operations only take place within the same node as the calling virtual machine, whereas write operations are synchronously performed on both nodes. This cross-node functionality is provided by the BlueBoxx layer and is transparent to the compute layer. Look at chapter 3.7 for a detailed description.

2.2 High-Performance

Common virtualization systems for high-performance and high-availability delegate the storage to special purpose appliances. This design increases the latency between storage appliances and compute nodes when compared to internal storage.

The configuration of hardware in both the storage and compute layers within the BlueBoxx can easily be customized to meet the customer's needs. Not only Business Intelligence and data warehouses but also applications with a high system load, such as OLTP applications, e-Commerce or web shop installations, benefit from BlueBoxx functionality and design.

2.3 Ready to Use

BlueBoxx is ready to use out-of-the-box. The installed software is configured and ready for the deployment of virtual machines and applications. All that's needed is to plug in the cables and boot it up.

For even faster and simpler deployment of applications the BlueBoxx provides several templates out-of-the-box:

- Database
- Application Server
- Webserver
- BI solutions

Through the use of a central administration interface as well as the pre-configured templates for the easy deployment of virtual machines the amount of effort required for administration is greatly reduced.

2.4 Cost-effective

The use of the BlueBoxx ensures a reduced ecological and economic footprint by following the principles of green IT. The BlueBoxx provides savings in operational expenses, which are achieved by:

- Minimizing space requirements
- Reduced power consumption
- Use of commodity hardware
- Open source software stack maintained through regular updates
- No need for external hardware components
- No dependency on hardware or appliance vendors with regards to compatibility or licensing

To summarize, the BlueBoxx is a pre-configured, ready to use product without the need for external components. It is a multi-purpose system designed to provide a highly available virtualization platform for any x86 operating system out-of-the-box. The BlueBoxx does not require a NAS or SAN infrastructure and is not bound to special hardware vendors or third party license fees.

3 BLUEBOXX ARCHITECTURE

This chapter describes the architecture of the BlueBoxx. We begin with an overview of the virtualization technology and then move on to an explanation of snapshots before closing with a description of the update procedure. The last chapter provides a complete list of references and sources.

3.1 Architecture Overview

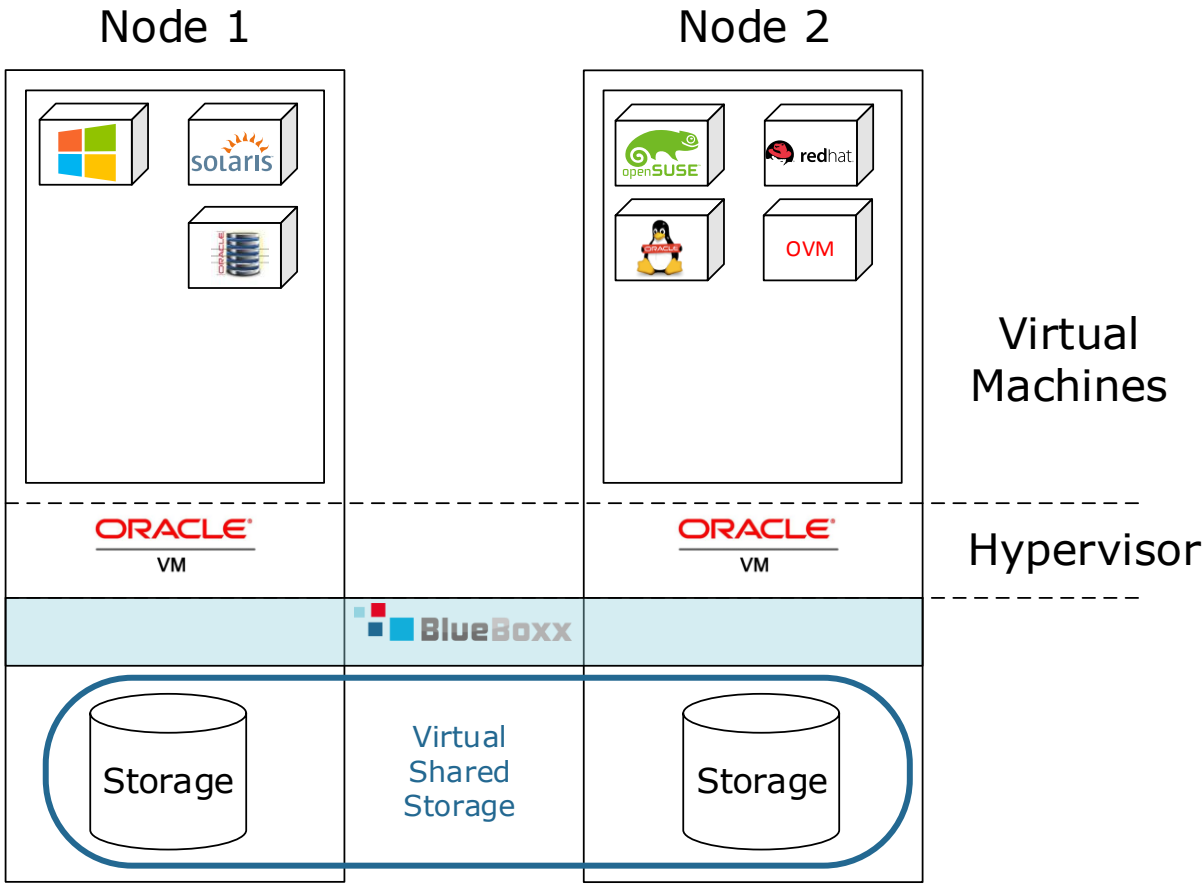

Figure 3 - Architecture Overview

- **Hypervisor:** Oracle VM is built upon the latest version of the Xen open source virtualization technology. Xen is a type 1 hypervisor running on top of the hardware on each node. The Xen architecture provides a single management domain, "dom0", and several user domains, "domU". In order to simplify administration Oracle VM ships with the Oracle VM Manager. This centralizes the administration of the management domain as well as providing hard partitioning to guarantee hardware resources to virtual machines. The Oracle VM Manager provides interfaces for automation, configuration and management. Please see the Oracle VM Concepts Guide [1] for a detailed explanation of Oracle VM Manager as well as the architecture of Oracle VM.
- **Virtual Machines:** Multiple operating systems can be installed, deployed and operated individually. This functionality is provided by the Hypervisor which is also able to run virtual machines in different virtualization modes. Chapter 2.7.1 of the Oracle VM Concepts Guide [1] describes the available modes.

The operating systems supported by Oracle VM include:

- Most common Linux based systems such as:
  - Oracle Enterprise Linux
  - Red Hat Enterprise
  - Suse Linux Enterprise Server
  - Oracle Solaris x86
- Microsoft Windows Server 2008
- Microsoft Windows Server 2012
- Microsoft Windows Server 2003

For a detailed certification matrix of the supported operating systems see the Oracle VM Release Notes [2].

3.2 Network Architecture

The BlueBoxx network has built-in redundancy to achieve high-availability and to guarantee fail-safe operation. The BlueBoxx network architecture defines three different networks, as shown in Figure 4.

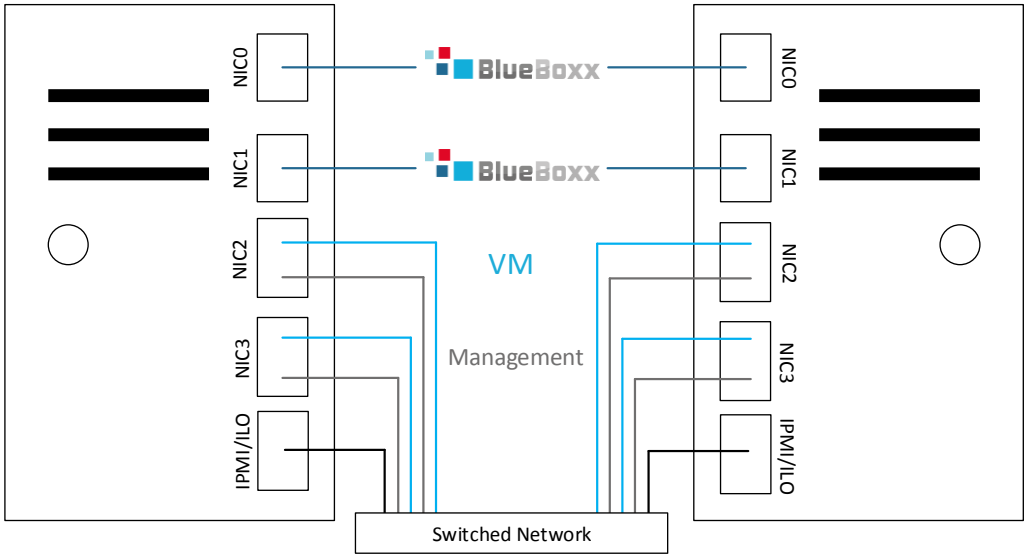

Figure 4 - Network Architecture

- BlueBoxx interconnect

  The BlueBoxx interconnect is used to enable synchronization, live migration and the high-availability of the two nodes within the BlueBoxx. This interconnect requires either a 10 Gbit/s switched network or a direct connection with a maximum cable length of 100 meters. The links in the BlueBoxx interconnect must be connected to different network cards on the same server to guarantee high availability and to ensure redundancy. The interconnect uses load balancing and therefore provides a bandwidth of 20 Gbit/s.

- **VM** network

  The VM network is used by virtual machines for internal and external communication and therefore is designed to provide client access to the BlueBoxx. The links in this network are operated in active-passive mode. This means that link one is active while link two is in standby mode ready to take over in the case of unexpected outage. It is recommended to use at least a 1 Gbit/s connection for the VM network.

- **Management** network

  The management network provides management and administration interfaces. This network is only accessible via the intranet and is meant for administration purposes only. The management network interfaces are operated in active-passive mode as described for the virtual machine network. It is recommended to use at least a 1 Gbit/s connection.

To configure the BlueBoxx the customer must provide the following information in spreadsheet-form, as shown in Figure 5:

- **Hostnames**

  The hostnames for each network node. These names can be chosen to meet the customer's network policy requirements.

- **V-LAN Tags**

  The virtual LAN tag for the management and VM network. This information is required if the customer operates a configured V-LAN infrastructure and can be defined to meet the customer policies.

- **IP address, Subnet Mask & Gateway**

  The IP address, subnet mask and gateway for each network node. This information is required for the routing and configuration of the network and can be defined to meet customer policies.

- **DNS Servers**

  The DNS servers as defined by customer policies or by existing customer infrastructure.

- **NTP Servers**

  For time synchronization it is recommended to operate at least two independent ntp servers. In environments where data is constantly shared between systems it is important that time is properly synchronized. It is therefore required that the ntp servers be reachable from the management network.

Device	Hostname	Interface	Description	VLAN Tag	IP address	Subnetmask	Gateway	DNS1	DNS2
Node1	nd1-tls-cl2	bond0	local node IP / management LAN	99	192.168.1.1	255.255.0.0	192.168.1.254	192.168.1.253	192.168.1.254
Node2	nd2-tls-cl2	bond0	local node IP / management LAN	99	192.168.1.2	255.255.0.0	192.168.1.254	192.168.1.253	192.168.1.254
Node1	nd1-tls-cl2	bond2	VM LAN (accessable by clients)	98					
Node2	nd2-tls-cl2	bond2	VM LAN (accessable by clients)	98					
Node1	il1-tls-cl2	ILO/Drac/IPMI	ILO/Drac/IPMI Interface (must be in the same range as bond0)	99	192.168.1.3	255.255.0.0	192.168.1.254	192.168.1.253	192.168.1.254
Node2	il2-tls-cl2	ILO/Drac/IPMI	ILO/Drac/IPMI Interface (must be in the same range as bond0)	99	192.168.1.4	255.255.0.0	192.168.1.254	192.168.1.253	192.168.1.254
Cluster-/ PoolName	cl2-tlc-cl2	n/a	management cluster IP (must be in the same range as bond0)	99	192.168.1.5	255.255.0.0	192.168.1.254	192.168.1.253	192.168.1.254
VM Manager/ Monitoring	mgr-tlc-cl2	n/a	OVM Manager (must reachable via bond0)	99	192.168.1.6	255.255.0.0	192.168.1.254	192.168.1.253	192.168.1.254
External Storage		n/a	must be in the same range as bond0	99		255.255.0.0	192.168.1.254	192.168.1.253	192.168.1.254
External Storage		ILO/Drac/IPMI	must be in the same range as bond0	99		255.255.0.0	192.168.1.254	192.168.1.253	192.168.1.254

NTP Servers reachable from VM Manager	
Server1	192.168.1.254
Server2	192.168.1.253
Server3	192.168.1.252
Server4	192.168.1.251

Please fill in if you want to connect a NFS/Iscsi Storage
Required

Figure 5 - Network configuration

3.3 High Availability functionality provided by the BlueBoxx

Heartbeat

Oracle VM uses OCFS2 as a base layer for virtual disks in a cluster configuration. OCFS2 ships with native cluster and high availability features which are extensively used by Oracle VM for live migration, synchronization and the detection of unexpected node outage within the configured server pool.

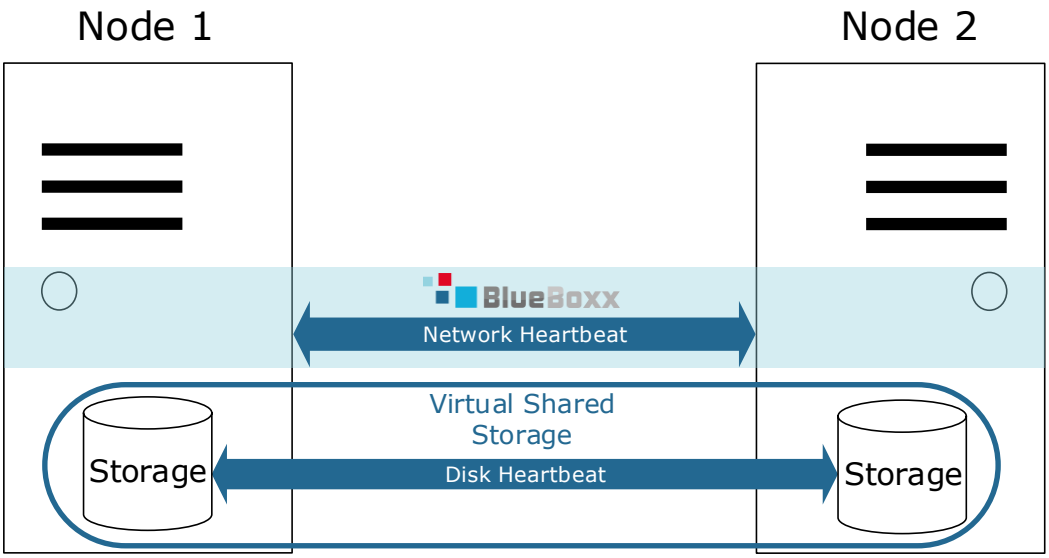

Figure 6 - Different Heartbeat Channels

OCFS2 uses multiple channels to detect absent nodes, as shown in Figure 6:

- **Disk Heartbeat**

  OCFS2 creates a region on the formatted disks reserved for the disk heartbeat. Each Oracle VM Server has access to OCFS2 storage and writes a timestamp to the reserved region. The VM Servers read the timestamps from other nodes to identify the active nodes in the cluster.

- **Network Heartbeat**

  The network heartbeat uses a TCP connection to send heartbeat packets at regular intervals to other nodes. The heartbeat is sent over the BlueBoxx interconnect.

Live Migration

The BlueBoxx provides live migration features to transfer virtual machines to another node without interruption of service. The interface to migrate a specific virtual machine is provided centrally by the Oracle VM Manager as shown in Figure 7.

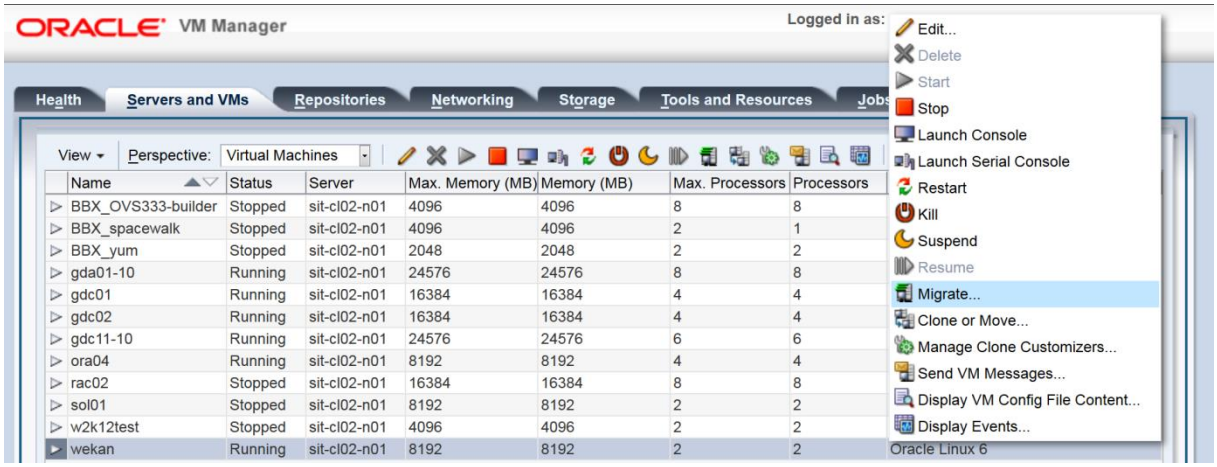

Figure 7 - Live migration in Oracle VM Manager

The live migration is based on two features of Oracle VM. These are:

- **Distributed Resource Scheduling**

  Distributed Resource Scheduling or DRS is responsible for maintaining the resources available within the Oracle VM cluster. If a defined CPU threshold is reached, Oracle VM live migrates the virtual machines to another Oracle VM Server without service interruption.

- **Distributed Power Management**

  Distributed Power Management or DPM is responsible for maintaining the utilization of the Oracle VM Server. Virtual machines on underutilized Oracle VM Servers are live migrated to other Oracle VM Servers. When all virtual machines have been migrated the Oracle VM Server will be shut down to reduce power consumption.

See also [1] for a detailed explanation of these features.

Cold Failover

The BlueBoxx provides all the high availability features of Oracle VM. A cold failover may be triggered automatically if a node is unreachable, or manually by the Oracle VM Manager. The Oracle VM architecture allows for a virtual-machine- specific configuration of high availability. Only virtual machines marked as HA are restarted in the case of a node outage. An automatic cold failover is triggered when a node is unreachable for 30 seconds. The remaining node then promotes itself to master status and restarts those virtual machines marked as HA from the unreachable node.

Example given:

- Node 1 is not reachable due to an unexpected outage

All the virtual machines deployed on node 1 are marked as HA, and have to be online as soon as possible. Node 2 reacts when the heartbeat of node 1 is absent, and the virtual machines from the unreachable node 1 are restarted on node 2, as shown in Figure 8.

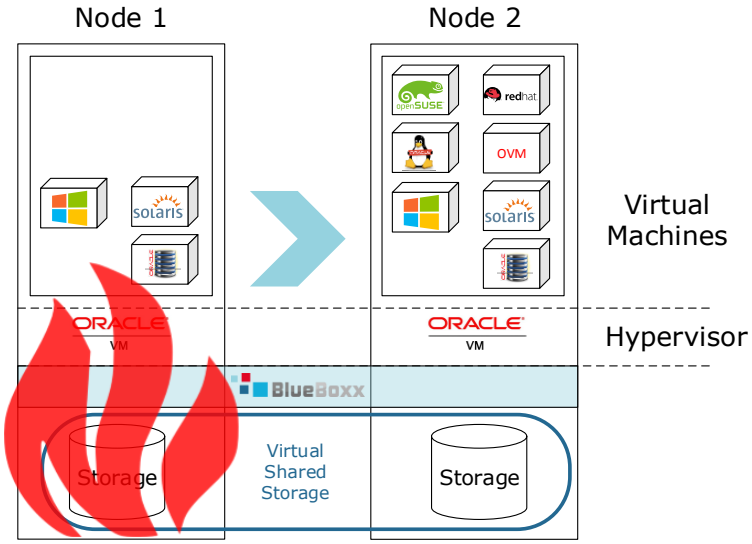

Figure 8 - Example: Node 1 is not reachable

3.4 Data integrity

The split-brain issue is a well-known risk in high availability systems. In order to avoid data corruption in the case of a loss of network connectivity additional fencing functionality has been integrated into the BlueBoxx.

Data consistency is the number one priority in case of an unexpected network outage. This is achieved by reacting flexibly in the case of a service interruption due to a server communication error. The monitoring system informs the administrator whenever a network link is unexpectedly lost or when unexpected behavior occurs. There are two different types of network interruptions:

1. **Single network link interruption**
   Through the use of redundant network connections the disconnection of a single link does not affect the operation of the BlueBoxx.

2. **Complete network interruption**
   In case of a complete network interruption three scenarios must be discussed:

   Complete interruption of the VM network

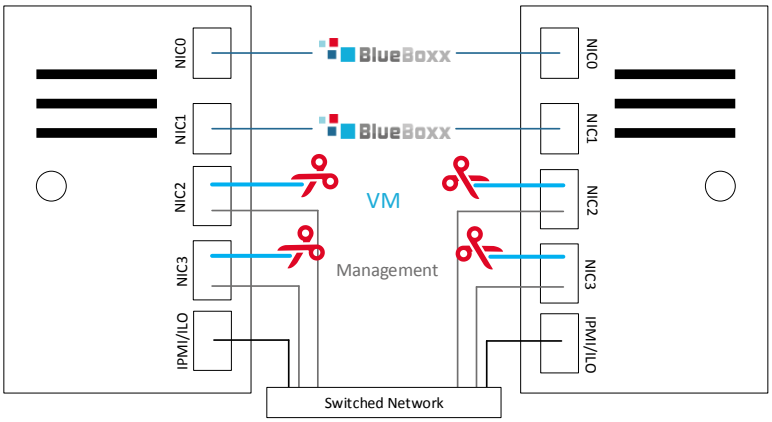

Figure 9 - Complete interruption of the VM network

The virtual machines are not reachable. The state of the virtual machines, Oracle VM and the storage is stable. The integrity of data is ensured and the BlueBoxx monitoring system informs the administrator about the network connection failure.

Complete interruption of the Management network

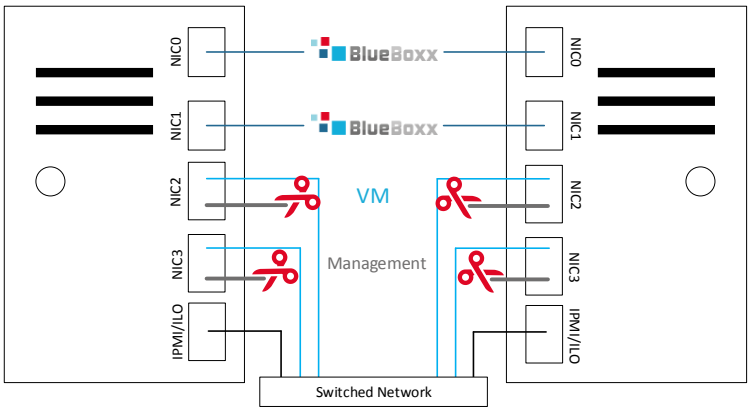

Figure 10 - Complete interruption of the Management network

When the management network is not accessible or available to a node the management and administration interfaces are no longer accessible. The virtual machines, as well as the BlueBoxx interconnect, are in a stable state and data integrity is ensured.

Complete interruption of the BlueBoxx interconnect

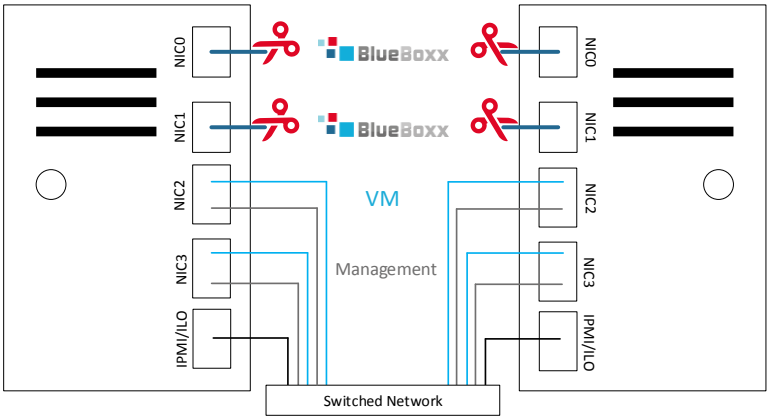

Figure 11 – Completely interruption of the BlueBoxx interconnect

If the BlueBoxx interconnect on node 1 is completely disconnected IO is immediately frozen and the management network is checked:

a) If the management network links are completely disconnected then node 1 is shut down to ensure data integrity.

b) If node 2 is reachable via the management network then both nodes are shut down to ensure data integrity.

c) If the management network is reachable from node 1 and node 2 is not reachable via the management network then node 1 is promoted to master and boots the virtual machines on node 1. The state of node 2 remains unclear and if the node still has access to the management network it will also be promoted. This is the only situation which would lead to a split-brain.

To avoid the split-brain situation it is highly recommended to operate the management network and BlueBoxx interconnect on separate physical networks. This ensures a connection between the nodes in the case of an unexpected outage or connection interruption on one of the networks.

3.5 High Performance

The BlueBoxx interconnect uses a 20 GBit/s connection to synchronize the virtual shared storage. It is possible to use the interconnect as a high speed connection between two virtual machines operated on separate nodes in the BlueBoxx. This is due to the low bandwidth consumed by the data synchronization. In order to use the interconnect it is required to configure a VLAN and attach it to the virtual machines.

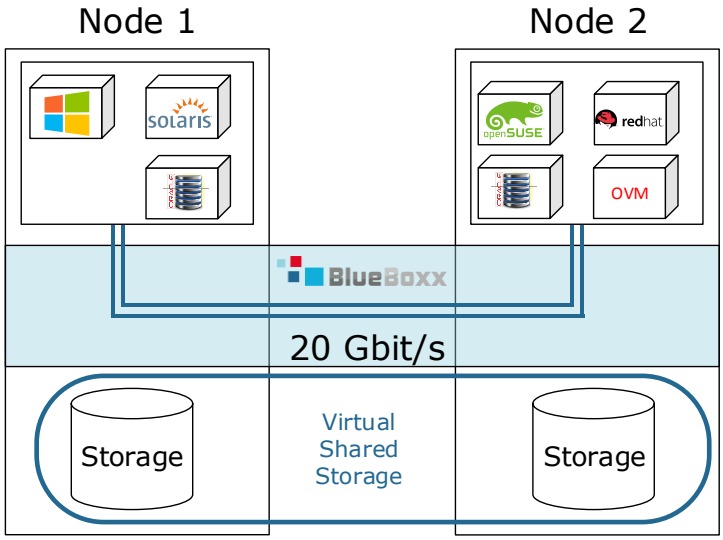

Figure 12 - Communication between two virtual machines over the BlueBoxx interconnect

3.6 Monitoring

The BlueBoxx ships with a centralized monitoring system to periodically check on the health of the BlueBoxx. This system provides checks on each node, some of which include:

- **CPU load**
  Displays the current and average CPU load.
- **RAID health**
  This sensor monitors the health of the RAID array and provides information about the status of the drives as well as the configured RAID level.
- **Network checks**
  Each network connection is monitored and its health displayed.
- **NTP status**
  The status of time synchronization to an NTP server including the status of the time offset to other nodes.
- **Dom0 Memory check**
  The relative and absolute consumed and available memory of dom0 is displayed.

Host	Service	Status
DemoManager	Current Load	OK
	Current Users	OK
	PING	OK
	Root Partition	WARNING
	SSH	OK
	Swap Usage	OK
	Total Processes	OK
sit-cl02-n01	BBX Loops	OK
	CPU Load	OK
	Dom0 Memory	OK
	MemoryProvisioning	CRITICAL
	Multipath	OK
	NTP sync	OK
	Network-Mgmt	OK
	Network-VMs	OK
	Network-VSS	OK
	OVS-Node1-State	OK
	Raid Status	OK
	VirtualStorage	OK
	VirtualStorageSyncProcess	OK
sit-cl02-n02	BBX Loops	OK
	CPU Load	OK
	Dom0 Memory	OK
	MemoryProvisioning	CRITICAL
	Multipath	OK
	NTP sync	OK
	Network-Mgmt	OK
	Network-VMs	OK
	Network-VSS	OK
	OVS-Node2-State	OK
	Raid Status	OK
	VirtualStorage	OK
	VirtualStorageSyncProcess	OK

Figure 13 - BlueBoxx monitoring

The following checks are BlueBoxx specific:

- **Provisioning check**

  The provisioning check provides information about a possible cold failover, as Oracle VM does not support over-provisioning. When one node fails the remaining node requires a certain amount of memory to restart the virtual machines from the failed node. A critical state, as shown in Figure 13, indicates that, in the case of a node failure, all high-availability virtual machines may not be restarted as the BlueBoxx would be unable to allocate enough memory.

- **Virtual Storage health**

  This check monitors the synchronization state of the virtual shared storage. If a node is rebooted synchronization will be stopped and the check turns yellow to signal a synchronization issue. As soon as the node is up and running again the synchronization process will be triggered and the VirtualStorageSyncProcess check displays the progress of synchronization. If the BlueBoxx interconnect is completely interrupted this check turns red.

3.7 Virtual Shared Storage

The BlueBoxx layer separates the organization of the storage layer from the compute layer and transparently mirrors all write operations to both storage nodes. The choice of hardware or hardware configuration has no effect on the operation of the virtual shared storage.

As shown in Figure 14, the write commit is sent to the virtual machine only once the data has been written to both nodes. Due to mirroring the optimal write performance depends on the speed of the BlueBoxx interconnect. A switched infrastructure may increase latency compared to the zero latency available when the nodes are directly connected.

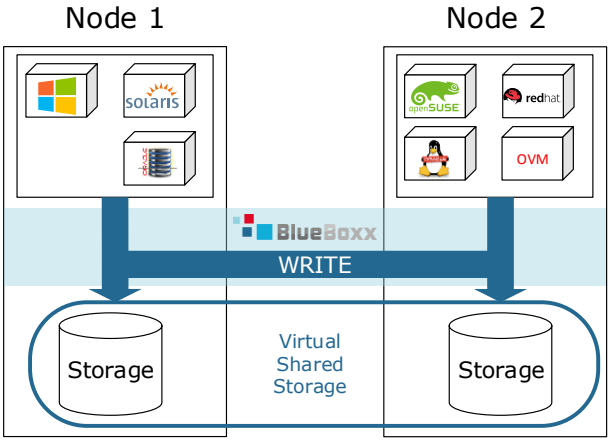

Figure 14 - Mirrored Writes

As shown in Figure 15 read operations only take place on the same node as the VM, and do not use the BlueBoxx interconnect.

If the storage in a node is unavailable due to maintenance the read and write operations of that particular node are redirected through the BlueBoxx layer to the other node, as shown in Figure 16. Benchmarks show that latency and performance are not significantly influenced by the redirection of read or write operations. Such a state is immediately visible in the monitoring systems as a warning message in the VirtualStorage check, indicating an inconsistent storage state.

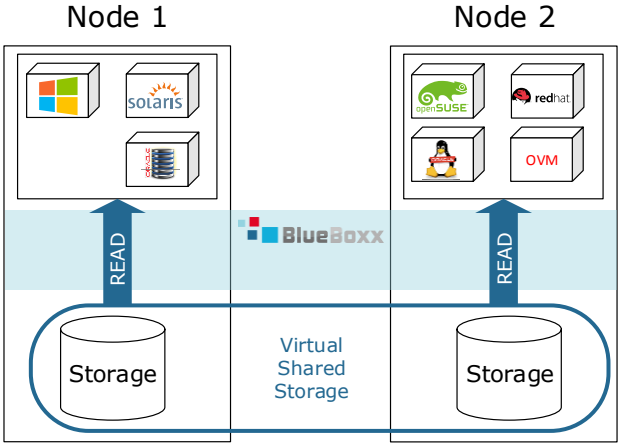

Figure 15 – Local Reads

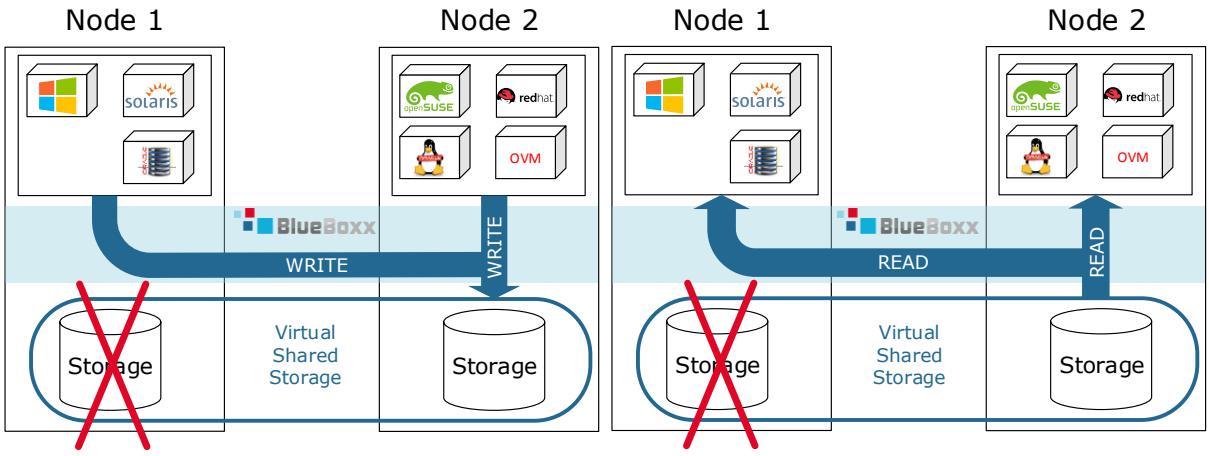

Figure 16 - redirected Read and Write Operations

The synchronization process will automatically be triggered as soon as the storage is available on node 1. The monitoring system displays the progress of the synchronization process in the VirtualStorageSyncProc check. Until the synchronization process has been successfully completed all read and write operations are still redirected to node 2 storage.

3.8 Backup

In a virtualized environment backups are taken using cloning and snapshot technologies. These are an inherent feature of virtual environments, and do not require any special purpose appliances.

The BlueBoxx utilizes full clones and copy-on-write clones to backup virtual machines to a backup target. The backup target does not need to be a special purpose appliance and therefore supports the BlueBoxx's cost-saving philosophy.

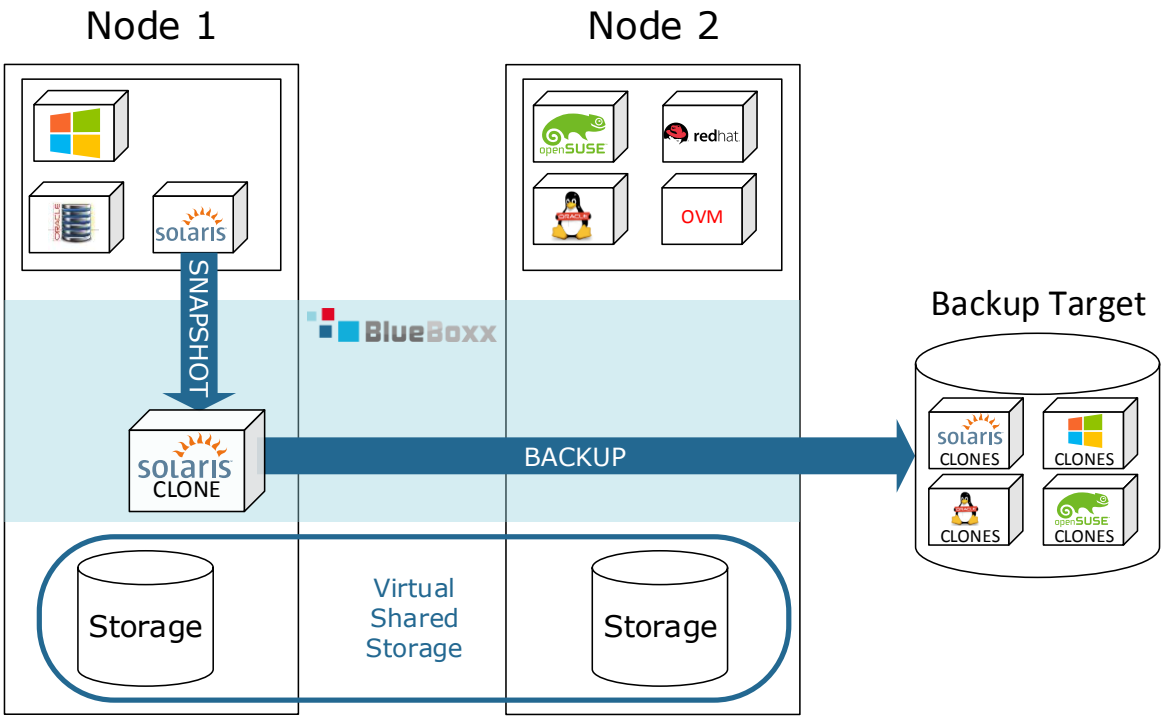

Figure 17 - Snapshots & Backup

A backup target must support one of the following communication options:

- Shared Network Attached Storage – NFS
- Shared iSCSI SANs
- Fibre Channel SANs
- Fibre Channel over Ethernet SANs

The BlueBoxx provides a backup system for hot and cold backups. Both methods are centrally managed by the Oracle VM Manager. To create a cold backup the virtual machine must be shut down, with the backup manually saved either on the same or a separate BlueBoxx, or to a connected SAN or NAS. There are two ways of performing a cold backup. The first option is a full backup, which copies the virtual disk file to a backup target. The second option is a Copy-On-Write snapshot, based on OCFS2 features.

The BlueBoxx provides integrated tag-based snapshot functionality to automatically create Copy-On-Write clones of running virtual machines. Created tags may be attached to a virtual machine in order to clone it. Tags are created in the Oracle VM Manager with the syntax of a backup tag as follows:

BBXBackup_<HH>:<MM>_<DOW>_<C2k>_<TargetRepository>_<Disk>

- **HH:MM**

  The time in 24 hour format for example: 22:45 is quarter to eleven pm

- **DOW**

  The day of week when the backup may be created. The range 0–6 where Sunday=0, Saturday=6 and every day=*

- **C2k**

  The number of clones to keep. After the given number the oldest clones will be dropped.

- **TargetRepository**

  The target storage repository configured within Oracle VM Manager. It is recommended to use a second BlueBoxx or external storage as the target repository.

- **Disk**

  The name suffix of the virtual disk defined in the Oracle VM Manager. The suffix is the last character combination. For example: the name of the disk is "sti05_d07_DATA" only the last character combination after the last underline identifies the disk. In this case the name is "DATA".

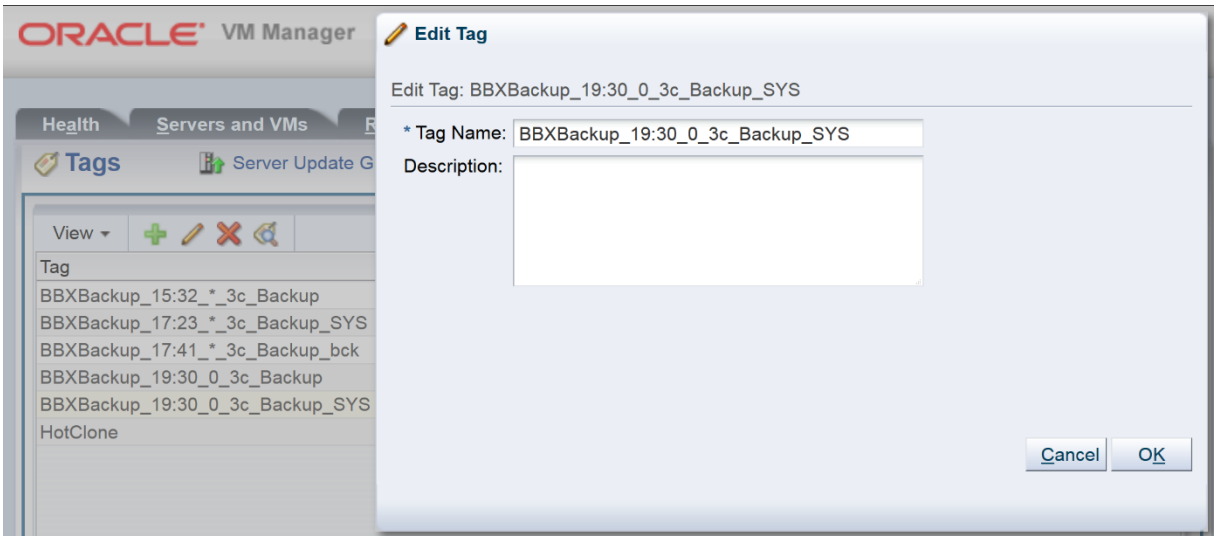

Figure 18 - Example Backup tag

An example of a tag is shown in Figure 18. This tag creates a hot backup of the virtual machines' "SYS" disk every Sunday 7:30 pm and keeps three clones. The name of the target repository is Backup.

With this mechanism no additional software is required, but we still strongly recommend performing file system backups as well as backing up the contents of the virtual machines.

3.9 Updates

Updates are provided centrally from a repository operated by Sphinx IT Consulting GmbH. Each BlueBoxx component can be updated separately to enable and support the planning of maintenance windows.

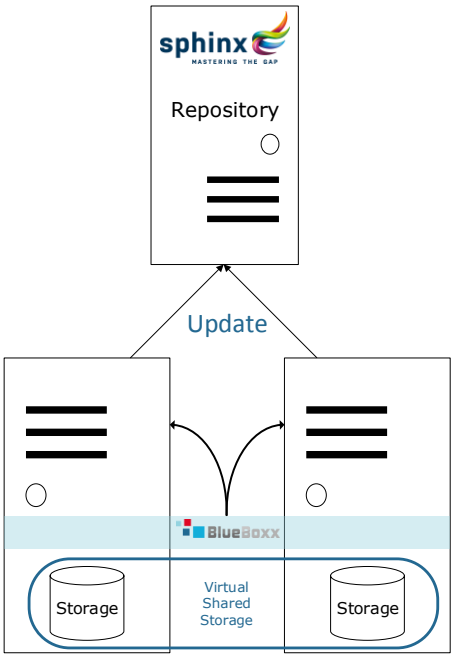

Figure 19 - Centralized update repository

All packages are built and maintained by Sphinx IT Consulting GmbH with updates centrally managed in the BlueBoxx administration and maintenance interface. Sphinx IT Consulting GmbH provides and maintains the following upgrade stacks based upon the original Oracle VM code distribution:

- **BlueBoxx**

  The BlueBoxx update maintains the BlueBoxx layer and takes effect after a node has been restarted. Due to live migration it is not required to schedule any downtime. This update is completely independent of other packages and may be performed by itself.

- **Xen**

  The Xen update maintains the installed hypervisor and takes effect after a node has been restarted. The Xen update does not depend on any other update and may be performed by itself.

- **Oracle VM Manager**

  Oracle VM consists of a manager and several components operated on the installed Oracle VM Server. The update of a manager does not require downtime, immediately taking place upon the reboot of the Oracle VM Manager and is independent of other updates.

- **Oracle VM Components**

  Due to dependencies it is recommended to update the Oracle VM components directly after a manager update. The component update requires a live migration and a restart of the node.

4 HARDWARE

This chapter describes the hardware requirements and the design of the network components required by the BlueBoxx.

4.1 Hardware Requirements

The BlueBoxx requires multiple networks and separates them by usage, as shown in Figure 20:

- The BlueBoxx interconnect is the backbone of the BlueBoxx layer.
- The local loops are used for the physical separation of the compute and storage layer.
- The VM network provides internal and external communication for the virtual machines.
- The Management network provides access to the BlueBoxx administration and monitoring services, and is used by the BlueBoxx high availability features.

See chapter 3 for a detailed explanation of the different networks and a description of the high availability functionality.

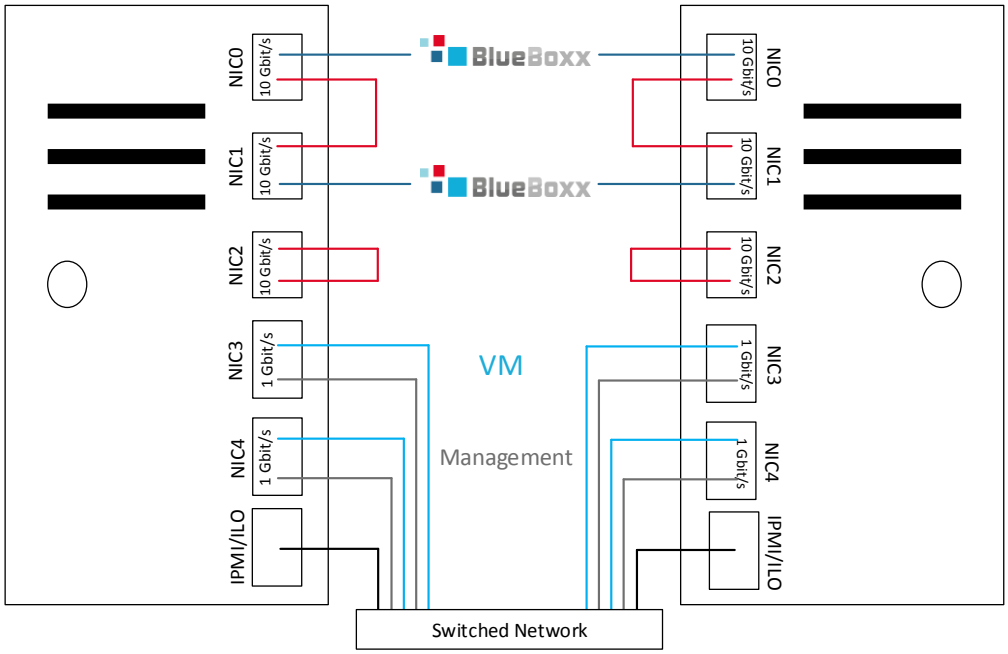

Figure 20 - BlueBoxx network architecture

The following are the hardware requirements for each node to guarantee high availability operation:

- RHEL 6 certified hardware. The use of alternate hardware would require certification by Sphinx IT Consulting GmbH
- Three Intel X540 AT1 and AT2 Dual Port 10 Gbit/s NICs for the BlueBoxx interconnect and local communication
- IPMI/ILO support to ensure monitoring services
- Two Dual Port 1 Gbit/s NICs for the VM and Management networks

It is recommended to use the latest processors from the Xeon family. Approximately Two CPU cores and 10 GB of RAM should be reserved for the Hypervisor, dependent upon expected load. For a high bandwidth connection to the virtual machines a 10 Gbit/s connection to a separate switched network is recommended. Individual hardware configurations must be certified by Sphinx IT Consulting GmbH.

4.2 Separation of BlueBoxx nodes

Dependent upon customer requirements the BlueBoxx may be operated in:

- **Same rack**

  The BlueBoxx nodes may be placed in the same rack with the same restrictions regarding the technical and physical environment.

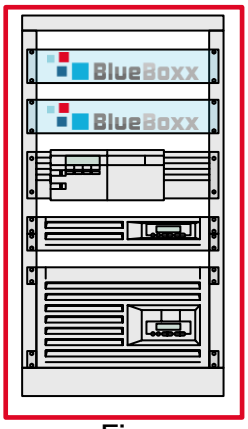

Figure 21 - BlueBoxx operation in a single rack

- **Two racks**

  The BlueBoxx nodes may be operated in two different racks within the same fire compartment and therefore with the same restrictions regarding the physical environment.

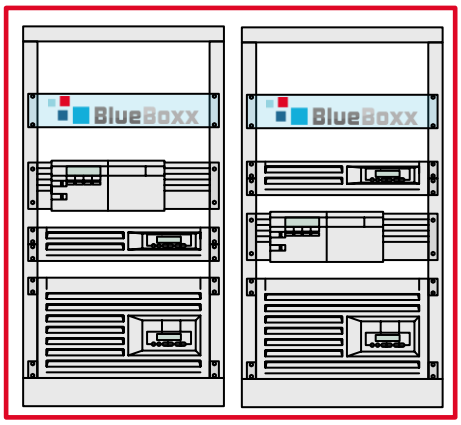

Figure 22 - Two rack operation

- **Separate fire compartments**

  The BlueBoxx may be operated on two independent locations and connected either directly or via a switched network. Due to the mirroring concept the switched network must be a 10 Gbit/s bandwidth low latency network.

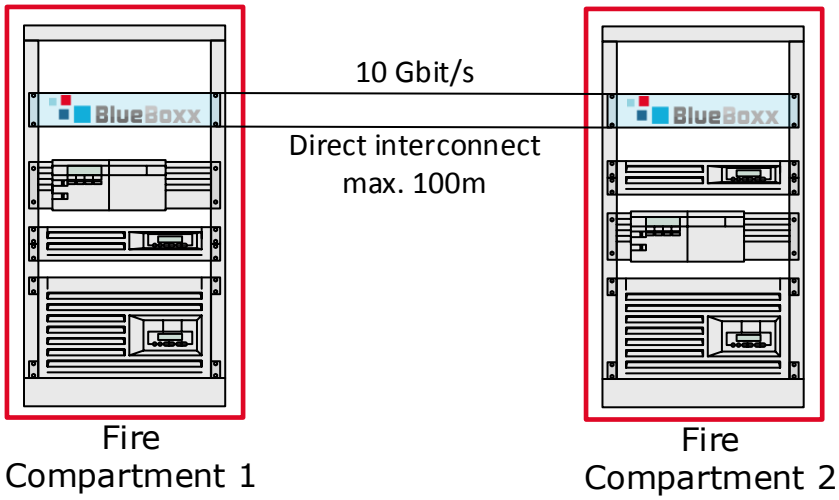

Figure 23 - Separate fire compartments, directly connected

The independent fire compartments allow the BlueBoxx to be highly available in the case of emergency and guarantee failsafe operation of the customers' services.

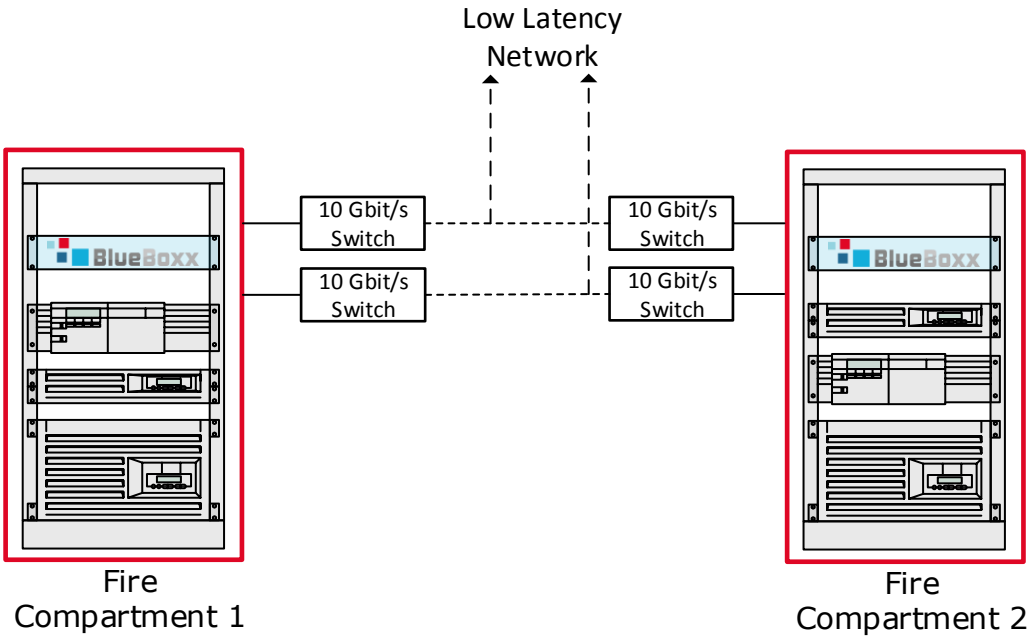

Figure 24 - BlueBoxx networking: Connection through switched network

The switched network may connect different buildings or may span a WAN connection. Due to the design a low latency connection is required to provide a 10 Gbit/s bandwidth.

5 BENEFITS FOR ORACLE CUSTOMERS

The BlueBoxx is based on Oracle VM. Oracle licensing rules and certification matrix are therefore applicable without restriction.

5.1 Support and Certification

The BlueBoxx is completely supported by Sphinx IT Consulting GmbH without the need for additional licenses or support from Oracle for Oracle VM on Oracle Enterprise Linux. Other Oracle products like Database or Fusion Middleware products are not part of the BlueBoxx and therefore must be licensed separately by the customer.

5.2 Licensing Options for Oracle Products

Oracle VM completely conforms to Oracle's hard partitioning license requirements. With Oracle VM it is possible to save licensing expenses on Oracle products. In contrast to other virtualization technologies it is not required to license the complete cluster where Oracle products are operated. As Oracle VM only requires you to license for a specific server or CPU core. Therefore, the following partitioning options are available for Oracle products:

- **Server Pinning**

  To operate an Oracle product on Oracle VM it is possible to pin the virtual machine to the server where the product is used. Therefore, only the CPUs of this node, defined in [3], must be licensed.

- **CPU Pinning**

  Oracle VM, and therefore BlueBoxx, provides the possibility to pin a specific number of CPU threads (vCPUs) to a virtual machine. This allows to license only these cores in contrast to licensing the complete cluster or node. CPU Core pinning completely conforms to the Oracle hard partitioning license requirements.

According to the document "Oracle Hard Partitioning" live migration of virtual machines with hard partitioning features enabled is not permitted. For a full explanation of the Oracle license and hard partitioning requirements see [3], [4] and [5].

6 GLOSSARY

HA	High-Availability
HP	High-Performance
OVM or OVM Manager	Oracle VM Manager
Dom0	Domain zero, management domain of the Xen hypervisor
DomU	User Domains, guests or deployed virtual machines of the Xen Hypervisor
NIC	Network Interface Card
DPM	Distributed Power Management
DRS	Distributed Resource Scheduler

7 REFERENCES

[1] Oracle Corporation, "Oracle VM: Concepts Guide for Release 3.3," 20 Juli 2015. [Online]. Available: http://docs.oracle.com/cd/E50245_01/E50249/E50249.pdf. [Accessed 21 März 2016].

[2] Oracle Corporation, "Oracle® VM: Release Notes for 3.3.1 Supported Operating Systems," Juli 2015. [Online]. Available: http://docs.oracle.com/cd/E50245_01/E50246/html/vmrns-guest-os.html. [Accessed 21 März 2016].

[3] Oracle Corporation, "Oracle© Software Investment Guide," 12 February 2016. [Online]. Available: http://www.oracle.com/us/corporate/pricing/sig-070616.pdf. [Accessed 20 April 2016].

[4] Oracle Corporation, "Oracle Hard Partitioning With Oracle VM x86," November 2013. [Online]. Available: http://www.oracle.com/technetwork/server-storage/vm/ovm-hardpart-168217.pdf. [Accessed 12 April 2016].

[5] Oracle Corporation, "Oracle Partitioning Policy," 6 April 2016. [Online]. Available: http://www.oracle.com/us/corporate/pricing/partitioning-070609.pdf. [Accessed 12 April 2016].

8 INDEX OF FIGURES

Figure 1 - The BlueBoxx Concept 5
Figure 2 - The two hardware layers of redundancy 6
Figure 3 - Architecture Overview 8
Figure 4 - Network Architecture 9
Figure 5 - Network configuration 11
Figure 6 - Different Heartbeat Channels 12
Figure 7 - Live migration in Oracle VM Manager 13
Figure 8 - Example: Node 1 is not reachable 14
Figure 9 - Complete interruption of the VM network 14
Figure 10 - Complete interruption of the Management network 15
Figure 11 – Completely interruption of the BlueBoxx interconnect 15
Figure 12 - Communication between two virtual machines over the BlueBoxx interconnect 16
Figure 13 - BlueBoxx monitoring 16
Figure 14 - Mirrored Writes 18
Figure 15 – Local Reads 18
Figure 16 - redirected Read and Write Operations 18
Figure 17 - Snapshots & Backup 19
Figure 18 - Example Backup tag 20
Figure 19 - Centralized update repository 21
Figure 20 - BlueBoxx network architecture 22
Figure 21 - BlueBoxx operation in a single rack 23
Figure 22 - Two rack operation 23
Figure 23 - Separate fire compartments, directly connected 24
Figure 24 - BlueBoxx networking: Connection through switched network 24