euroNAS HA Cluster

Best Practice for Running VMware

on

euroNAS SAN and HA Cluster

Rev. 14-07-12

euroNAS believes that the information in this publication is accurate of its publication date. The information in this publication is provided “as it is”. euroNAS GmbH makes no warranties of any kind with respect to the information in this publication. euroNAS GmbH specifically disclaims implied warranties or fitness for a particular purpose.

Index

Introduction

This document describes how to use and build a highly-available storage that ensures business continuity with VMware ESX even in the event of total storage server failure.

It will also explain you the main differences and advantages/disadvantages of euroNAS TCP/IP based HA Cluster and SAN Cluster that uses fibre channel protocol.

Until recently, building a highly available, performance fibre channel storage that provides up to 99.99% availability was very expensive. euroNAS changes this by providing a powerful, scalable high-availability solution that is both affordable and easy to deploy. It brings to small and midsize companies the enterprise functionality, usually affordable only to large companies.

What makes euroNAS Cluster different from other solutions?

On the market there are many SAN solutions providing high speeds and redundancy. They will protect you from disk failures or failure of one RAID controller.

euroNAS goes a step further by protecting you from a total failure of one of the units and also enabling you to simultaneously have the same data on 2 different places.

This will protect your data in case of some catastrophic events like fire, earthquake or theft and provide you business continuity at the same time.

Also not many manufacturers on the market will offer you both types of server mirror (TCP/IP and Fibre Channel) enabling you to use the same solution provider depending on your budget and performance requirements.

Advantage of euroNAS SAN software not being locked to specific hardware

euroNAS runs on standard x64 servers. This enables you total freedom in choosing the hardware that fits to your performance, scalability and budget.

Legacy storage manufacturers often use proprietary hardware and an operating system which will only run on the hardware in question. This makes it difficult to expand in the future and forces you to buy new units. euroNAS will grow with the needs of your company. It is not locked to proprietary hardware – this makes it a highly flexible and scalable solution.

By selecting euroNAS storage software you are not only getting highest possible redundancy in case of server failure but also hardware tolerant operating system that runs on any hardware.

Which euroNAS Cluster is the best for me?

Both, euroNAS HA Cluster and SAN Cluster will provide you with the reliability, scalability and simple management. However depending on the protocol in use or system load you can experience different consistency in performance.

The best answer would be – it depends on your performance requirements, existing infrastructure and budget.

Difference between Ethernet (iSCSI & NFS) and Fibre Channel

iSCSI and NFS are protocols that use TCP for transport and enable you to use the existing TCP/IP network infrastructure. iSCSI has been developed as cost efficient alternative to fibre channel. It is built on an underlying TCP/IP protocol and is usually implemented only as a software initiator that incurs significant processing overhead. Similar to Fibre Channel, iSCSI will present its targets to iSCSI initiators directly as block devices. NFS will present the device at the file system level.

Greatest benefit of iSCSI is the cost of the solution. In comparison to FC, iSCSI requires less expensive hardware and since it is based on Ethernet, more IT Administrators are familiar with the technology. However, Ethernet is not designed for transferring block data in a networked storage environment.

One of the greatest problems of the Ethernet is its way of handling data collisions – when more than one computer tries to transmit data simultaneously. Under heavy load condition, too many packet collisions will greatly reduce the whole network efficiency due to retransmissions. When an iSCSI or NFS path is overloaded, the TCP/IP protocol drops packets and requires them to be resent. FC has a built-in pause mechanism when overload occurs. On iSCSI and NFS this can lead to oversubscribed network paths and the performance degrades more and more because the dropped packets need to be resent.

On iSCSI and NFS this problem can be reduced by lowering the traffic, bonding the network ports or using switches with larger port buffer.

Fibre channel with its asynchronous protocol design ensures that even under heaviest load, collisions are handled efficiently and maintains maximum throughput.

While fibre channel constantly performs very close to its max speed (8 Gb = 800 MB/s or 16 GB = 1600 MB/s), collision management on Ethernet makes it very difficult to achieve same consistent performance.

Benefit of euroNAS Cluster

Higher availability

When using standard Storage your network applications depend on this single point of failure.

Whereas many appliances will provide you with redundant controllers and components – most of them will not protect you from a total unit failure.

When using euroNAS Cluster your network applications will continue to work - no matter which server fails, clients can still access their data. Even with proper backup policy it takes a while to recover and manual intervention to get server and data back online.

With euroNAS Cluster, should a hardware or software fault occur, it is detected through intelligent features and the software automatically moves storage requests to the other server mirrored in the cluster.

Server Mirror

euroNAS advanced technology creates a cluster of two mirrored servers with real time, continuous data replication and synchronization. Both servers contain identical data securing redundancy in case of server failure.

Simple Management interface

euroNAS Cluster is perfect for IT Professionals. You are not only gaining a high available storage solution with continuous replication but also an efficient and simple to use management interface. Whole server and cluster configuration can be done within 10 minutes.

euroNAS Cluster components and features

SAN Cluster

In order to achieve high-availability SAN Cluster consists of following parts

Cluster node

Cluster node represents an individual member server of the cluster.

Minimum requirements are 4 GB RAM and at least 2 QLogic Fibre Channel HBA ports. One is used for internal replication, other for providing FC targets to the clients.

Active node

Active node is the server that is providing storage to the clients. It replicates data to the passive node and makes sure that data is consistent on both nodes

Passive node

Passive node contains the same data as the active node and stays in standby mode as long as the other node is active. In case of failure of an active mode, it will automatically take over and continue to provide storage to the clients

Cluster drive

Cluster drive represents a mirror of 2 drives on each individual server.

This 2 drives can be a single disk or raid array. They are mirrored in real-time. On this drives Fibre Channel Targets are installed.

You can have up to 50 cluster drives.

Corporate connection

Corporate connection is the FC connection used by clients for accessing fibre channel cluster targets

Replication connection

Internal fibre channel connection used for data replication between the nodes

ALUA (Asymmetric Logical Unit Access)

ALUA support delivers high availability by establishing multiple sessions from a client to each node of SAN Cluster through Fibre Channel. In the event that a device in the path fails, I/O requests will be automatically redirected to an alternate path for continued application availability.

On euroNAS SAN Cluster the active node will be automatically recognized as active and the passive node will be automatically recognized as standby node by the client.

HA Cluster

In order to achieve high-availability HA Cluster consists of following parts

Cluster node

Cluster node represents individual server that is member of the cluster. All nodes are equal – from each node you can monitor and configure cluster services.

Cluster drive

Cluster drive represents a mirror of 2 drives on each individual server.

This 2 drives can be a single disk, software or raid array. They are mirrored in real-time. On this drives shares and iSCSI Targets are installed.

This drives don't necessary need to use full size of the available space. It is possible to create multiple cluster drives on the same disk or RAID array. They can be then split across the servers.

For example:

On each individual server there is a RAID array defined of 16 TB. On cluster drive creation you can create 2 separate cluster drives of 8 TB size. Each 8 TB cluster drive can be defined to run on one of the servers. All resources defined on this cluster drive (iSCSI, NFS, AFP/CIFS) will run on this server. This way the load between the servers will be balanced and you will achieve best possible performance.

Cluster resource

Cluster resource represents network share or iSCSI Target within the cluster. This can be SMB/CIFS & AFP, NFS share or iSCSI Target. Each resource has its own size and is accessible through the ip address defined for the cluster drive.

Corporate network

Corporate network is the network used by network clients for accessing cluster resources

Replication network

Internal network used for heartbeat and data replication between the servers

Cluster Access IP Address

Cluster Access IP Address is the address that enables clients to access shares and iSCSI Targets. The greatest advantage is that this address is shared by both HA cluster systems. If any of these servers fail – the cluster IP address will move to the running server.

- Cluster Shares and iSCSI Targets will not be available for client connections until you create an Access IP Address for the Cluster Drive
- Access IP Address can only point to one cluster drive
- The IP Address must be in the Corporate Network

Cluster access IP address is defined per cluster drive. All resources on this drive are available through this IP address.

Network Test IP

The purpose of this IP(s) is to prevent storage desychronization, also known as “split brain” situations. Such situation may occur when both nodes are running, but one or both are disconnected from the network.

HA Cluster nodes will ping this address constantly and check if still is able to reach the network. It is recommended to define more than just one IP address.

IP address on any pingable device on your network that is constantly on, router, gateway, printer, mail server etc. can be used as network test IP.

In one of the nodes is unable to ping this IP it will automatically know that it cannot reach the network and push all services to other node.

If a node cannot reach all defined test IP addresses it will shut down services to make sure that data remains consistent.

Cluster Events

Within cluster there are several types of events that can occur. Depending on type transfer of services can take different time. Initial time for servers to realize that other server has failed is around 30-40 seconds. After this resources are moved pretty quickly.

SAN Cluster

Estimated time to move around 10 resources on another server is around 10-20 seconds. To move around 30 resources usually takes about 50 seconds.

HA Cluster

Initial time for servers to realize that other server has failed is around 30-40 seconds. After this resources are moved pretty quickly.

Estimated time to move around 10 resources on another server is around 50-60 seconds. To move around 30 resources usually takes about 70-80 seconds.

These times may vary depending on the performance of the hardware and the size of the storages.

Failover

Failover occurs when server detects server fault and automatically moves all resource on another server. On failover services are moved faster on another server. One example would be total crash of one of the servers.

Switchover

Switchover happens when access to a resource is manually moved from one server to another. This is usually done if you wish to perform system maintenance of one of the nodes. Moving services takes little bit longer than on failover.

Hardware recommendations

Installation Disk

For best performance the euroNAS cluster software should be installed on an SSD with a minimum size of 8 GB. Installation disk is separate from data.

Data Disk

Since installation disk is exclusively used for the OS, for data at least one additional disk is needed. Data disk can be simple disk or disks managed by the hardware RAID controller.

System RAM

The server should have at least 8 GB of RAM

Fibre Chanel HBA (SAN Cluster only)

Per node it at least 2 port QLogic HBA is required. One port will be used for replication, other for presenting FC targets to the initiators.

Networking (SAN Cluster only)

Networking is not used for replication but it is necessary for management and internal communication between the nodes.

There must be one or more reliable IP Addresses on the Corporate Network that both servers can reach for proper operation of the High-Availability/Custer features

Since it is only used for internal communication and management – 1 Gigabit network is sufficient.

DHCP should not be used unless static leases are provided

Using DHCP without static leases will cause the cluster to lose communication to other node and fail on IP address change

Networking (HA Cluster)

There must be one or more reliable IP Addresses on the Corporate Network that both servers can reach for proper operation of the High-Availability/Custer features

An example would be the router which is on the Corporate Network or connects the servers to the Corporate Network

For best performance and high availability the server should have 4 individual Network Interface Cards with a minimum bandwidth of 1 Gigabit (10 Gigabit recommended). They should be configured as two separate bonded networks. One of the bonds should be connected to the Corporate Network. The other bond should be connected directly to the other server on a private Replication Network

If the servers are in the same location they should be cabled directly to each other without any switches between them if possible.

For top performance the Replication Network should be 10 Gigabit

If dual-port cards are used each port on the card should be assigned to a separate bond so that the failure of one card does not break the entire bond

Card 1 Port 1 -> bond0 (Corporate Network)
Card 1 Port 2 -> bond1 (Replication Network)
Card 2 Port 1 -> bond0 (Corporate Network)
Card 2 Port 2 -> bond1 (Replication Network)

Quad-port cards should be avoided

DHCP should not be used unless static leases are provided

Using DHCP without static leases will cause the cluster to fail on IP address change

Performance recommendations

The bottleneck of euroNAS Cluster is the disk I/O and the speed of the FC/Ethernet controller. We recommend the use of SAS or SSD disks and hardware RAID controller.

euroNAS Cluster is hardware independent however we still recommend using identical hardware. Due to server replication in real time, the performance will depend on the weakest link. Using slower disks on one server will slow the server with faster disks as well.

Limitations

SAN Cluster detects automatically many possible fail scenarios and moves resources automatically. There are some scenarios where SAN cluster is unable to detect the problem.

Replication network failure – corporate network available from both nodes (HA Cluster only)

In this case server can reach network test IP from both nodes so both nodes think that they are online and available. On the other hand, replication network is unavailable. In order to prevent possible data loss server replication and resources are halted until replication network becomes available again. In order to prevent this we recommend network card teaming (port failover)

Powering down master node during sync process

During the sync process the node with more current data must not be powered off. In this case server replication service will stop the drives in order to prevent data corruption.

Only server with less current data can be powered off. This does apply if both nodes are in sync. In this case either of the servers can be powered off or rebooted.

VMWare Configuration

Failover Concept

Both servers are proven and are working reliably in most demanding VMWare environments. However HA Cluster and SAN cluster differentiate in failover concept.

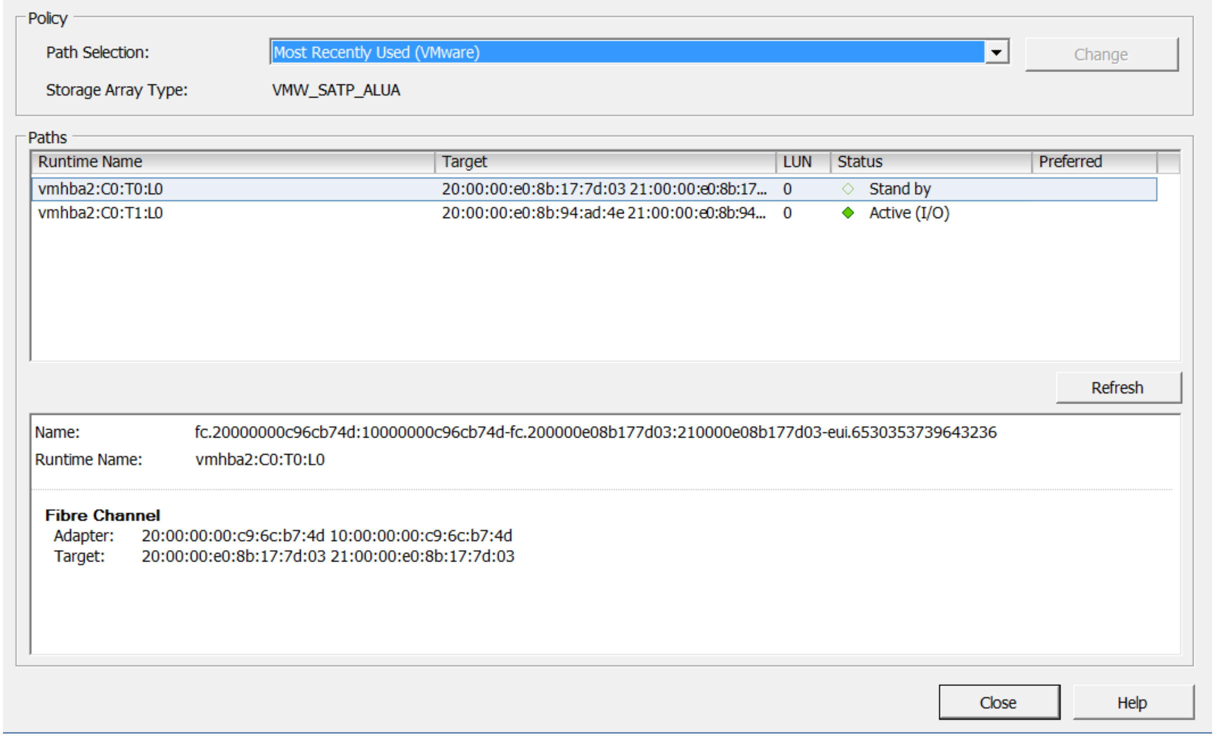
HA Cluster

HA Cluster will provide its resources using the virtual IP connection. If failover event occurs, it will move this virtual IP Address to the working node. VMWare ESX is a cluster aware OS and will wait for the failover to finish. VMs can experience small freeze during failover but will continue to work normally after that.

SAN Cluster

ESX Server supports multipathing, multipathing allows you to define more than one physical path that transfers the data. In case of failure, ESX will automatically switch to another working physical path.

SAN Cluster will communicate to the ESX Server using Asymmetric Logical Unit Access (ALUA) – it will notify VMware which node it should use and which one should be in standby. If failover event occurs, remaining port will automatically register itself as active, VMware will automatically continue to use this port. VMs can experience small freeze during failover but will continue to work normally after that. Usually the freeze is much shorter than on HA Cluster.



Policy

Path Selection: Most Recently Used (VMware)

Storage Array Type: VMW_SATP_ALUA

Paths

Runtime Name	Target	LUN	Status	Preferred
vmhba2:C0:T0:L0	20:00:00:e0:8b:17:7d:03 21:00:00:e0:8b:17...	0	◇ Stand by	
vmhba2:C0:T1:L0	20:00:00:e0:8b:94:ad:4e 21:00:00:e0:8b:94...	0	◆ Active (I/O)	

Name: fc.20000000c96cb74d:10000000c96cb74d-fc.200000e08b177d03:210000e08b177d03-eui.6530353739643236

Runtime Name: vmhba2:C0:T0:L0

Fibre Channel

Adapter: 20:00:00:00:c9:6c:b7:4d 10:00:00:00:c9:6c:b7:4d

Target: 20:00:00:e0:8b:17:7d:03 21:00:00:e0:8b:17:7d:03

Best Practice

HA Cluster

In a production environment using at least gigabit network is a must. For the best and most optimal performance we recommend 10 Gigabit connections.

Because of the latency issues of the Ethernet technology we recommend using HA Cluster within a LAN – for WAN configurations we recommend SAN Cluster.

It is recommended to separate LAN traffic between VMWare and iSCSI/NFS and keep it separate from other network traffic. One way of achieving that would be to configure the switch to use ports for communication into a separate VLAN.

Also VMkernel network can be configured to separate the iSCSI/NFS traffic separate from other networks, including the management and virtual machine networks.

Because iSCSI and NFS use TCP/IP to transfer I/O, you can expect some latency – to decrease the latency you should remove as many components that will cause hops between the storage and ESX server. We don't recommend to route traffic and suggest keeping both ESX and Storage on the same subnet.

In high-availability environment we recommend teaming network controllers for greater redundancy.

It is recommended that the Network card and switch support hardware based flow-control.

Advanced Settings

VMware will work "out of the box" – however if used with iSCSI and many iSCSI resources we recommend following settings to be changed in the Advanced Settings of the VMware ESX.

RecoveryTimeout – 120 sec

NoopTimeout – 30

SAN Cluster

It is recommended to separate FC Targets accessed by ESX from other devices. This way you can prevent non-ESX systems from accessing the targets and possibly destroying VMFS data. This can be done by using zoning. Zoning is configured on FC switch and defines which HBAs can connect to which targets. If configured, the devices outside a zone are not visible to the devices inside the zone. Within euroNAS SAN Cluster you can alternatively define which WWNs are allowed to access and see the targets.

VMware supports several different path selection policies, we recommend using the "Fixed AP" policy. ESX will always use the preferred path to the disk defined by euroNAS ALUA.