OUTDATED

StarWind Virtual SAN®
Reference Architecture for Dell PowerEdge R730

FEBRUARY 2015

VERSION 0.9

WHITE PAPER

Trademarks

Changes

Technical Support and Services

If you have questions about installing or using this software, check this and other documents first - you will find answers to most of your questions on the Technical Papers webpage or in StarWind Forum. If you need further assistance, please contact us.

Copyright ©2009-2015 StarWind Software Inc.

In 2016, Gartner named StarWind "Cool Vendor for Compute Platforms".

Gartner does not endorse any vendor, product or service depicted in its research publications, and does not advise technology users to select only those vendors with the highest ratings or other designation. Gartner research publications consist of the opinions of Gartner's research organization and should not be construed as statements of fact. Gartner disclaims all warranties, expressed or implied, with respect to this research, including any warranties of merchantability or fitness for a particular purpose.

About StarWind

StarWind is a pioneer in virtualization and a company that participated in the development of this technology from its earliest days. Now the company is among the leading vendors of software and hardware hyper-converged solutions. The company's core product is the years-proven StarWind Virtual SAN, which allows SMB and ROBO to benefit from cost-efficient hyperconverged IT infrastructure. Having earned a reputation of reliability, StarWind created a hardware product line and is actively tapping into hyperconverged and storage appliances market. In 2016, Gartner named StarWind "Cool Vendor for Compute Platforms" following the success and popularity of StarWind HyperConverged Appliance. StarWind partners with world-known companies: Microsoft, VMware, Veeam, Intel, Dell, Mellanox, Citrix, Western Digital, etc.

Contents

Introduction

This document is for StarWind users who want to deepen their knowledge of StarWind Virtual SAN and its benefits. It gives insight into our work in the virtualization sphere, achievements and the common configurations of building storage virtualization layer using Dell servers and StarWind Virtual SAN.

StarWind Software is a pioneer of storage virtualization. It has been on the market since late 2008 and was founded as a spin-off from Rocket Division Software – a company started in 2003. StarWind had been providing customers with a commercial storage-over-IP solutions long before iSCSI was invented, adopted and became mainstream. StarWind developed first iSCSI initiator for Windows (before IBM and Microsoft did) and one of the first commercial iSCSI targets for Windows. StarWind was the first company to develop a storage virtualization stack running directly on top of a hypervisor and not as a virtual machine, something done by VMware two years later for their vSphere hypervisor under "Virtual SAN" brand name. Basically StarWind had closed first million in sales of that product two years before VMware released their beta version. Right now StarWind Software continues to innovate, surfing on a bleeding edge of storage and virtualization technologies wave. Everything we do, invent, and implement years before others catch up is done to protect our customers' ROI. They can be sure that our product and concepts they adopted for production will become a mainstream soon and stay "hot" for a while.

StarWind Virtual SAN is entirely software-based, hypervisor-centric virtual machine storage. It creates a fully fault-tolerant and high-performing storage pool purpose-built for the virtualization workload. StarWind Virtual SAN basically "mirrors" inexpensive and fast internal storage between hosts using Ethernet. Internal storage can be high-capacity modern SATA spindles, server-grade SAS or uber-fast PCIe-attached flash. StarWind Virtual SAN completely eliminates any need for an expensive Fibre Channel, iSCSI SAN, various NFS/SMB NAS, or other physical shared storage like SAS JBODs typically required to unlock all virtualization potential. It seamlessly integrates into the hypervisor landscape as a combination of user-land services and kernel-mode drivers or as a virtual machine for unbeatable performance and exceptional simplicity of use.

Dell is a great company and probably number one server hardware manufacturer. It provides exceptionally good build quality, strict service level agreements, outstanding support and unbeatable pricing. It is here to stick for a while so a customer can rest assured their virtualization hardware vendor will not go out of the business tomorrow. "Rock-solid" in all terms is the reason why StarWind Software decided to build our Reference Architecture using Dell-branded hardware.

StarWind Virtual SAN can convert any x86 server into a SAN. It allows building of highly available hyper-converged setups, where hypervisor, virtual machines, and the SAN are running on the same physical server.

In this document we describe the server configuration and architecture specifically designed to be used with StarWind Virtual SAN. Along with: Dell PowerEdge R730 as the main hardware building block, Windows Server 2012 R2 as the base operating system for StarWind and Windows Server 2012 R2 Hyper-V as the hypervisor platform.

Mission

To provide the customer, who can actually be an end-user, a system integrator or an OEM builder, with a simple step-by-step cookbook on how to build a virtualization platform for different deployments and workloads: server virtualization or virtual desktop infrastructure. The idea is to squeeze all possible IOPS paying as little money as possible and at the same time without compromising cluster stability and construction quality.

Building blocks

Components for the project are both software titles (StarWind Virtual SAN, OS, hypervisors) and hardware (Dell servers, flash memory modules, hard disks, Ethernet switches, and so on). Some components are mandatory and are present in all provided diagrams. These are StarWind Virtual SAN and actual Dell servers. Other components are optional and can be missing from some sample configurations. VMware vSphere hypervisor is deployed only with VMware targeted installation and flash memory is not always used but only for IOPS-savvy deployments.

Software

StarWind Virtual SAN®

StarWind Virtual SAN is an entirely software-based, hypervisor-centric virtual machine storage. It creates a fully fault-tolerant and high-performing storage pool that is purpose-built for the virtualization workload from the ground up.

Virtual SAN is a natural part of the hypervisor. The result is both outstanding performance and unified administration using hypervisor-specific management tools. Virtual SAN "gets the job done" with all major virtualization platforms running on Hyper-V and Windows as a native application and on vSphere or Xen nested inside a VM. Non-virtualized clusters typically deployed for performance-intensive SQL Server and Exchange installations are also fully supported.

Virtual SAN provisions and manages storage with VM workload in mind. General-purpose storage arrays are usually the best candidates to serve hypervisors. Virtual SAN uses Log-Structured File System and sophisticated caching in order to avoid as much of a random I/O dominating VM world as possible but still without deploying extremely expensive and cost-inefficient all-flash model.

I/O latency is reduced and much of the network traffic is eliminated by utilizing Virtual SAN distributed RAM and flash-based caches.

Virtual SAN basically "mirrors" internal storage and caches between the given numbers of hosts creating a fault-tolerant storage pool. It is completely up to the system administrator as to how many replicas of a particular VM or LUN are kept alive and how many active storage controllers a cluster should use. Individual disks, memory modules, whole compute and storage hosts may fail but uptime is never compromised.

Virtual SAN comes with an effective mechanism to ensure that the mission-critical business data is replicated to a disaster recovery site. Replication is implemented to be asynchronous, background, deduplication-and compression-aware, as well as snapshot-based. Primary site I/O operations do not suffer from leaned resources and WAN channel requirements being sparse.

Virtual SAN flexibly adopts both Scale-Up and Scale-Out architectures. Capacity can be increased by simply throwing more spindles into the existing storage cluster node. Bringing in a new host with its own CPU, RAM and internal storage can scale storage capacity, I/O performance and VM number crunching facilities out.

StarWind Virtual SAN utilizes inexpensive commodity hardware. “Magic” happens with the help of proprietary in-house developed software and commodity servers, MLC flash, spinning disks, and Ethernet. Virtual SAN can definitely take care of higher-performing SAS, more reliable SLC flash, and a faster 10 or 40/56 GbE network but all of that “higher league” gear is completely optional.

Virtual SAN implements inter-node tiering technology to offload cold data such as snapshots from fast and expensive primary storage to slower but inexpensive secondary storage. The combined result allows of the use of a much smaller amount of flash as primary tier with cheap spindles as a secondary tier.

In addition to the inter-node tiering, StarWind Virtual SAN implements VM and flash-friendly space reduction technologies such as in-line deduplication and compression. The result allows of the increased usable capacity of all-flash configurations, boosts I/O performance as more data can be pinpointed to now virtually “bigger” deduplicated caches and also prolongs flash life since the amount of write operations is reduced dramatically.

Microsoft Windows Server 2012 R2

Since the release and presentation of a Hyper-V 3.0 Microsoft is now treated as a serious contender of the for virtualization market. It basically took away SMBs and ROBOs from VMware and now continues its advance into the Enterprise scale. Within our implementation Windows Server is used as a hypervisor (Hyper-V is just a role enabled in a Windows Server 2012 R2 installation) and also as a platform to run StarWind Virtual SAN in a so-called “bare metal” setup for “Compute and Storage Separated” installations covering for both Hyper-V and vSphere. Within these configurations StarWind Virtual SAN is running on a set of Windows-running Dell servers and provides “external” hypervisor cluster with a shared storage over iSCSI and SMB3 protocols for vSphere and Hyper-V.

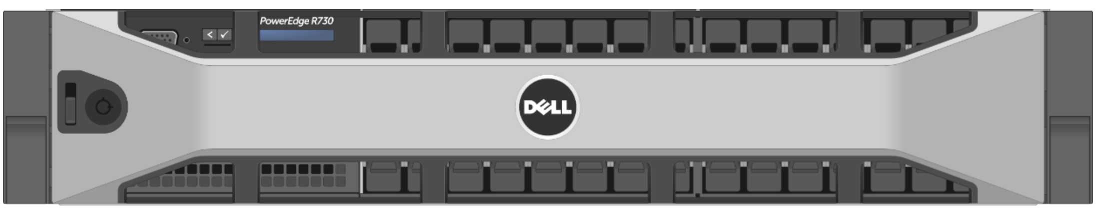
Microsoft has its own initiatives for Software Defined Storage (SDS) targeting both smaller and bigger virtualization deployments, and they are called Clustered Storage Spaces and Scale-Out File Server respectively. Clustered Storage Spaces are block- based and Scale-Out File Server is a SMB3-based file front-end or gateway on top of it to work around limited Clustered Storage Spaces scalability. Clustered Storage Spaces are not completely “software” as Microsoft heavily relies on an existing all-SAS storage infrastructure and thus needs SAS hard disks and / or flash (both high capacity SATA and uber-fast PCIe-attached flash are out of game) mounted into external SAS JBODs (no way to use internally-mounted server storage with Clustered Storage Spaces even if it is SAS). Additionally: SAS HBAs and cabling to create all-connected mesh between two or three servers and set of SAS JBODs. Inter-node data is wired with SAS fabric and distributed access locks are not present within Microsoft SDS implementation - they are using SAS hardware ones.

On the other hand, StarWind Virtual SAN does not need SAS connectivity, it wires inter-node communications over 10 or 40/56 gigabit Ethernet which is both faster and more cost effective. StarWind does not rely on SAS hardware locks so can use internally mounted SAS capacity and also inexpensive SATA hard disks and high performing PCIe flash. StarWind Virtual SAN adds some more important features like in-line write-targeted deduplication pairing Microsoft off-line read-targeted one, log-structured file system dramatically improving performance for all-random typical VM workload, and distributed write-back caches pairing Microsoft read-only CSV cache and write-back flash cache. So at the end of the day StarWind Virtual SAN provides Hyper-V nodes with a virtual shared storage with no SAS-related hardware dependencies in hyper-converged scenario or completely takes control of a block back-end being basement for a Scale-Out File Server in “Compute and Storage Separated” scenario.

Hardware

Dell PowerEdge R730

R730 is the best fit for the hyper-converged virtualization environment building block role. It provides a customer with the widest possible choice of components and also supports multiple internal storage combinations including all types of storage, from 3.5” disks and PCIe flash to 1.8” flash. The Intel® Xeon® E5 2600 v3 line provides a broad selection of CPUs to be used depending on the implementation scenario.



Technical specifications:

Processor

Up to 2 Intel® Xeon® E5 2600 v3 series processors.
The CPU ranges from 1.6 GHz 6 core / 6 threads Intel® Xeon® E5 2603 v3 to 2.3 GHz 18 core / 36 threads Intel® Xeon® E5 2699 v3

Chipset

Intel C610 series chipset

Memory

Up to 768GB (24 DIMM slots): 4GB/8GB/16GB/32GB DDR4 up to 2133MT/s

Storage

HDD: SAS, SATA, Near-line SAS SSD: SAS, SATA
16 x 2.5" – up to 29TB via 1.8TB hot-plug SAS hard drives"
8 x 3.5" – up to 48TB via 6TB hot-plug NL SAS hard drives"

Drive Bays

Internal hard drive bay and hot-plug backplane:
Up to 16 x 2.5" HDD: SAS, SATA, Near-line SAS SSD: SAS, SATA
Up to 8 x 3.5" HDD: SAS, SATA, Near-line SAS SSD: SAS, SATA

RAID Controllers

Internal:
PERC H330
PERC H730
PERC H730P
External:
PERC H830
Network Controller
4 x 1Gb, 2 x 1Gb + 2 x 10Gb, 4 x 10Gb

Typical server configurations

Configuration 1: SAS storage, hyper-converged
2 x Intel Xeon E5-2650v3 Processor
2.5" chassis with up to 16 Hard drives
8x 16 GB RDIMM RAM, 128 GB total
16x 600GB 15K RPM SAS 6Gbps 2.5in Hot-plug Hard Drive
PERC H730 integrated RAID controller, 1GB NV Cache
Broadcom 57800 2x10Gb DA/SFP+ + 2x1Gb BT Network Daughter Card (SAN + LAN)
Intel Ethernet I350 QP 1Gb Server Adapter (LAN)
iDRAC8 Enterprise w/ vFlash, 8GB SD
Dual, Hot-plug, Redundant Power Supply (1+1), 750W

Usable capacity:
up to 4.8 TB with SAS RAID 10
up to 9.6 TB with SAS RAID 0
Total: $ 18,427.83 [1,2]

Configuration 2: Hybrid storage, hyper-converged
2 x Intel Xeon E5-2650v3 Processor
3.5" chassis with up to 8 Hard drives
16x 16 GB RDIMM RAM, 256 GB total
4x 4 TB 7.2K RPM SATA 6Gbps disks
2x 800GB Read intensive Solid State Drive SAS MLC 12Gbps
PERC H730 integrated RAID controller, 1 GB NV Cache
Broadcom 57800 2x10Gb DA/SFP+ + 2x1Gb BT Network Daughter Card (SAN + LAN)
Intel Ethernet I350 QP 1Gb Server Adapter (LAN)
iDRAC8 Enterprise w/ vFlash, 8GB SD
Dual, Hot-plug, Redundant Power Supply (1+1), 750W

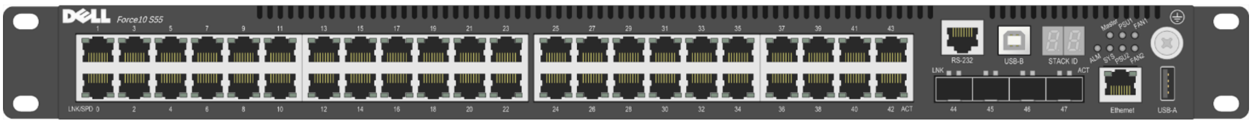
Usable capacity:
up to 8,8 TB with SATA RAID 10 & SAS SSD RAID 1
up to 12,8 TB with SATA RAID 5 & SAS SSD RAID 1,
up to 17,8 TB with SATA RAID 0 & SAS SSD RAID 0
Total: $ 16,980,92 [1,2]

Configuration 3: all-flash, hyper-converged
2 x Intel Xeon E5-2690v3 Processor
3.5" chassis with up to 8 Hard drives
24x 16 GB RDIMM RAM, 384 GB total
8x 800GB Mix use Solid State Drive SATA MLC 12Gbps
PERC H330 integrated RAID controller
Broadcom 57800 2x10Gb DA/SFP+ + 2x1Gb BT Network Daughter Card (LAN)
Mellanox ConnectX-3 Dual Port 40Gb Direct Attach/QSFP Server Network Adapter (SAN)
iDRAC7 Enterprise w/ vFlash, 8GB SD
Dual, Hot-plug, Redundant Power Supply (1+1), 750W

Usable capacity:
up to 3,2 TB with SAS SSD RAID 10
up to 5,6 TB with SAS SSD RAID 5
up to 6,4 TB with SAS SSD RAID 0
Total: $ 29,749.62 [1,2]

Ethernet switch

Dell S55 has been chosen as the main switch for all hypervisor and user traffic in the environment. With 44 1-GbE, and 4 10-GbE ports it can easily accommodate all the management and hypervisor traffic in the given configurations.



Additional 10 GbE switches like Dell N4810 may be added to support iSCSI backbone scale-out beyond 3 nodes.

[1] The prices are not final and provided for reference only.
Prices were calculated using the "Customize and buy" section on Dell web site.

[2] Default service contract selected.

StarWind Virtual SAN® Reference Architecture

StarWind Virtual SAN supports both hyper-converged and "Compute and Storage Separated" virtualization scenarios. With Hyper-Converged scenario both hypervisor and storage virtualization stack share the same hardware so most of I/O happens locally bypassing Ethernet wiring. With "Compute and Storage Separated" hypervisor cluster and storage cluster are physically stand-alone entities and all I/O happens over Ethernet fabric.

We will start with a setup built "from scratch" to match I/O and other requirements of customer setup. We will also leave some space for expansion and provide details on how to do Scale-Out of the cluster to match updated needs (more VMs, more served storage etc) later in the Scale-Out section.

Hyper-converged

StarWind Virtual SAN simplifies the virtualization environment architecture by unifying the storage and compute powers into a single layer making each server an equal part of the cluster. This results in a hyper-converged unit which does not require any dedicated shared storage resource. The minimum number of servers required for a fault tolerant configuration is 2. The minimum number of servers for a fault tolerant configuration with storage setup in RAID 0 is 3.

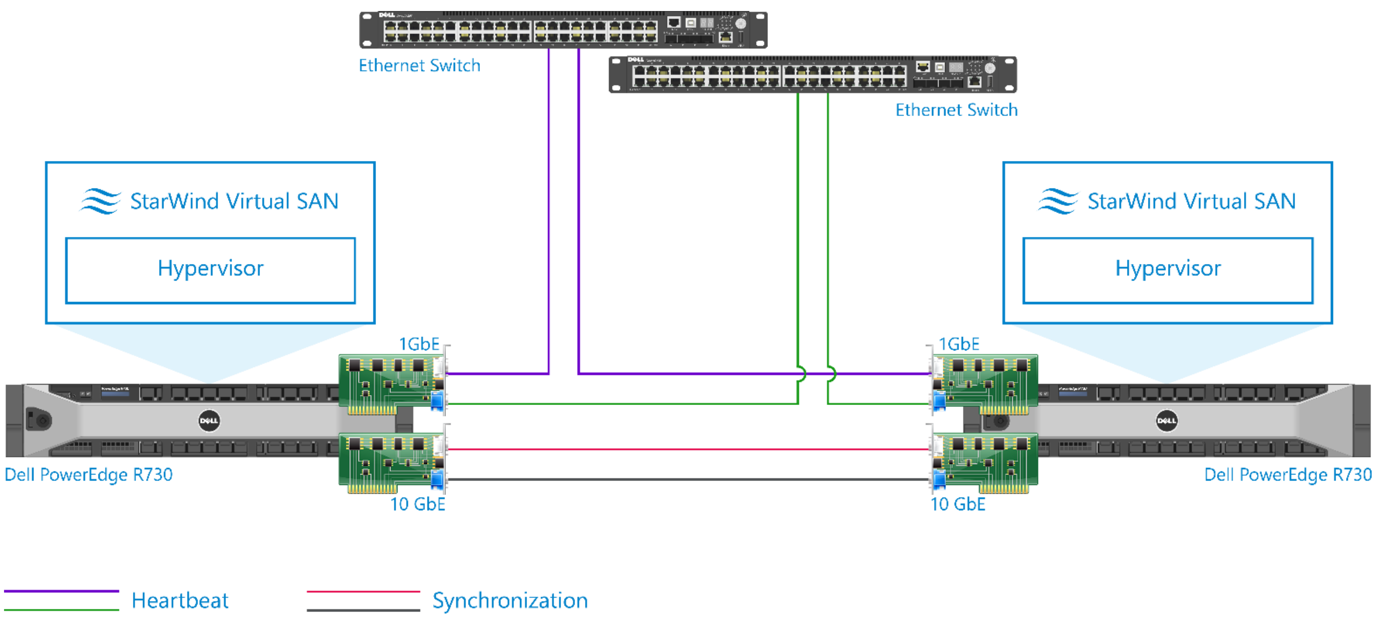


Fig. 1 – Hyper-converged 2-node setup, the diagram only shows the StarWind interconnections. Cluster networking should use separate networks

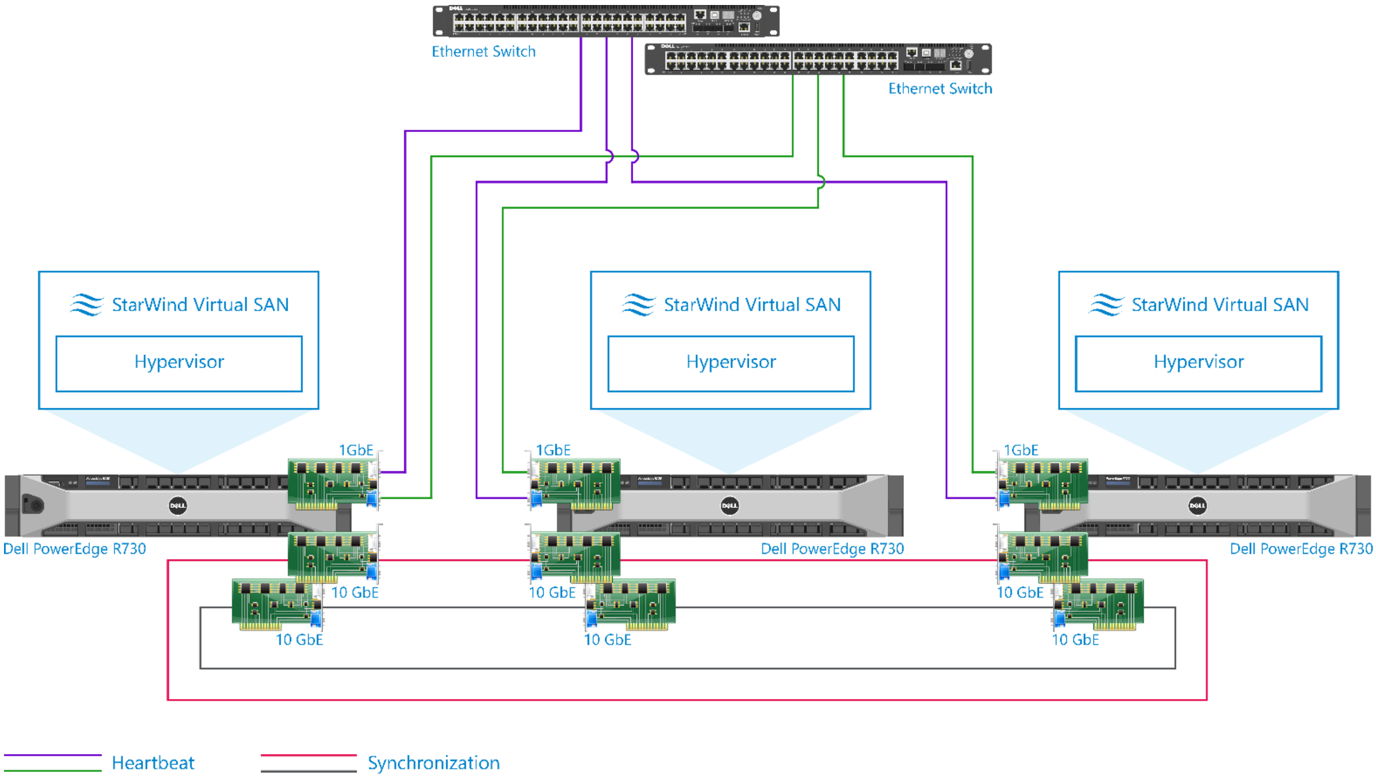


Fig. 2 – Hyper-converged 3-node setup, the diagram only shows StarWind interconnections.

All read I/O in this type of setups is local only (and can also be aggregated using partner nodes thanks to MPIO). Writes are only confirmed with the partner nodes mirroring the same LU so their amount is reduced, RAM and Flash cache is local so with data in cache there are no Ethernet sends or receives. Since server's RAM is used for the hypervisor, VM needs, and StarWind Virtual SAN L1 cache, sizing RAM is very important. Provisioning insufficient amount of RAM to the hypervisor, VMs, or StarWind can lead to a performance decrease of the corresponding component or whole configuration as such.

Memory configuration

Minimal possible setup starts with 16GB of RAM. We assume 4 GB are to be reserved for the hypervisor and base OS. StarWind Virtual SAN L1 cache memory requirement starts with 1 GB. Each VM typically requires 2-4 GB of RAM. Also, the server has to maintain some reserve RAM resources to accept VMs migrated from another cluster node. As a result we can provision no more than 40% of server's RAM for the locally hosted VMs. E.g. with 1 TB of clustered storage this configuration has about 11 GB of RAM available for clustered VMs.

Storage Configuration

Storage is described by 2 key characteristics: performance and capacity. In virtualization, main performance metric is the number of IO commands it can process per second (IOPS). This value has to be determined based on thorough application IO monitoring and calculation of multiple applications working together on one host as VMs. The number of IOPS underlying storage can process has to be at least equal to the sum of all VM IOPS together. Additional power reserve is then calculated and added to the storage requirement for the future performance growth.

Capacity determines how much data you can store on a given server without adding SAN & NAS external storage. The storage array of the server is chosen based on these 2 characteristics. This process is usually unique for each environment so there is no rule of thumb for storage array to be chosen depending on the company size or number of VMs.

Networking

The interconnections in the virtualization environment also determine how fast the users can communicate with the applications running inside virtualization environment.
The minimum networking configuration is 4x 1-GbE ports. In this configuration 2 ports are used exclusively for StarWind traffic. 1 Port is used for VM migration between the hosts, and another for the management, client, and StarWind heartbeat connections. In a 3-node hyper-converged setup additional 2 ports per node are required.

In a clustered hyper-converged environment, networking also determines the maximum write performance, because each written block has to be transferred to the partner cluster node to guarantee storage redundancy. Thus, the throughput of the links used for StarWind synchronization and iSCSI traffic has to be higher than the maximum required storage performance in order to avoid bottlenecks.

An additional synchronization and iSCSI connection has to be added as soon as storage traffic starts maxing out the existing interconnects.

Currently, one of the most cost effective ways to interconnect a 2-node hyper-converged setup is a direct 10 GbE connection between the nodes. It allows both throughput capabilities for iSCSI and cluster traffic, as well as maintains a power reserve to scale the configuration.
With a 3-node setup 10 GbE direct connections are also supported, but only for iSCSI and synchronization traffic.

Contacts

US Headquarters	EMEA and APAC
1-617-449-7717	+44 20 3769 1857 (UK)
1-617-507-5845	+49 302 1788 849 (Germany)
	+33 097 7197 857 (France)
	+7 495 975 94 39 (Russian Federation and CIS)
	1-866-790-2646

Customer Support Portal:	https://www.starwind.com/support
Support Forum:	https://www.starwind.com/forums
Sales:	sales@starwind.com
General Information:	info@starwind.com

StarWind

https://www.starwind.com **StarWind Software, Inc.** 35 Village Rd., Suite 100, Middleton, MA 01949 USA