

READY FOR **ANY** vForum2015

9 December 2015 | Taipei, Taiwan

Virtual SAN 關鍵應用實務分享

Hawk Tsao 曹惟超
Professional Service Consultant

Agenda

1 快速回顧 “Virtual SAN overview and what’s new”

2 設計

3 概念驗證

4 佈署, 監控

5 Take away

Virtual SAN Overview

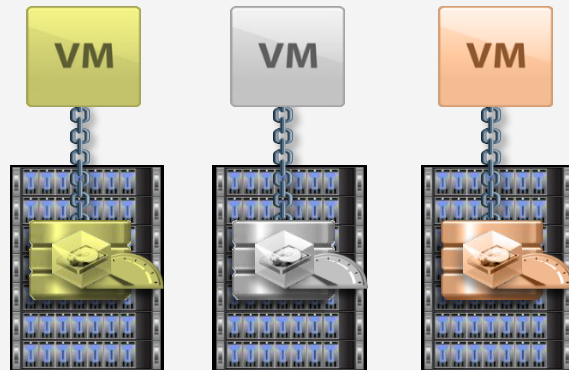
今日所面臨的多重挑戰

Specialized Expensive HW



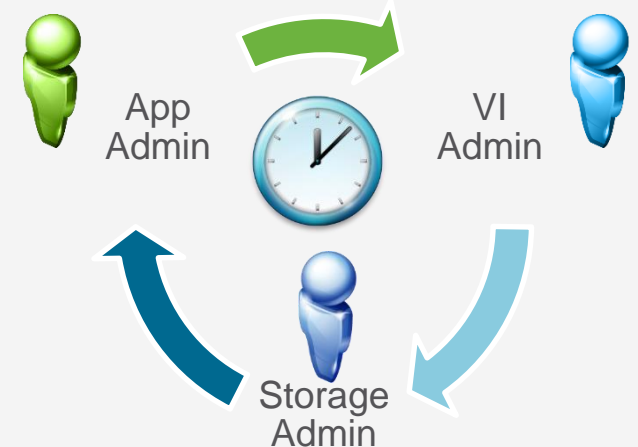
- ✗ Not commodity
- ✗ Low utilization
- ✗ Overprovisioning

Device-centric Silos



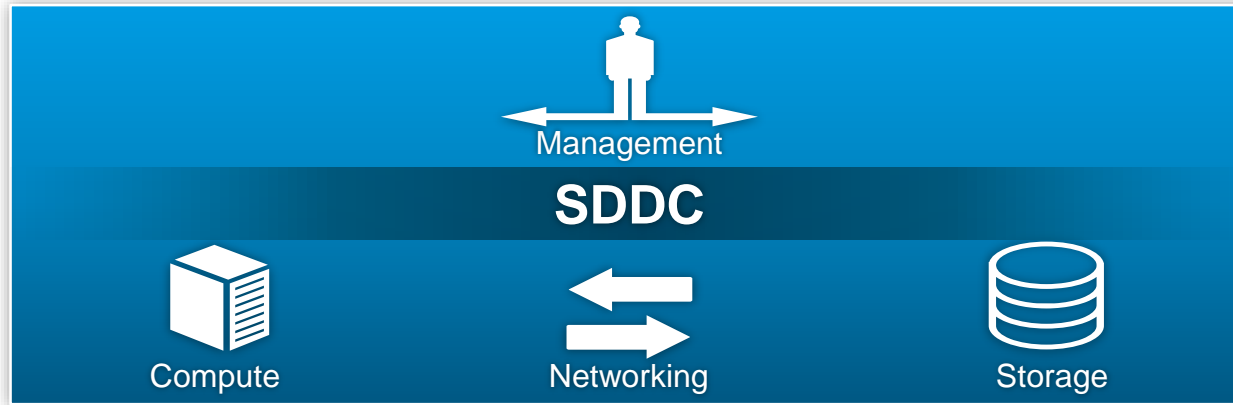
- ✗ Static classes of service
- ✗ Rigid provisioning
- ✗ Lack of granular control
- ✗ Frequent data migrations

Complex Processes



- ✗ Time consuming processes
- ✗ Lack of automation
- ✗ Slow reaction to request

超融合架構: The Ideal Architecture for SDDC



Open Systems /
Traditional Infrastructure



Hyper-Converged
Infrastructure

- ✓ Simplicity
- ✓ Cost
- ✓ Scalability
- ✓ Performance

彈性的資訊基礎建設選擇

更具彈性

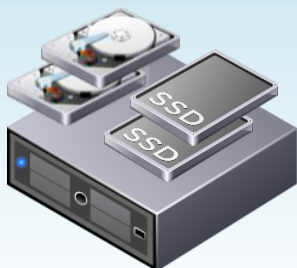


易於使用

Software + Hardware



Virtual SAN +
vSphere

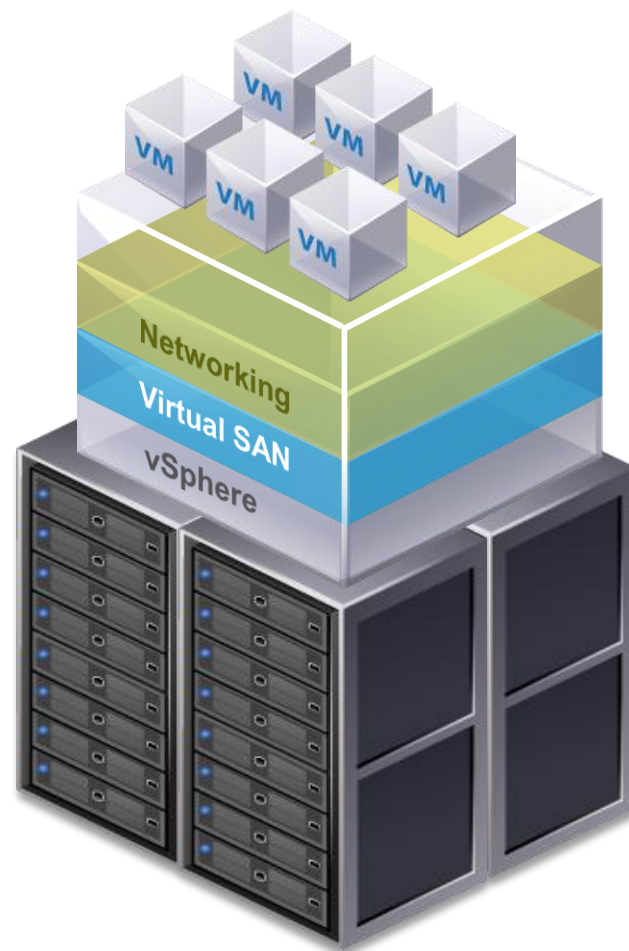


Virtual SAN
Ready Node

Integrated Systems

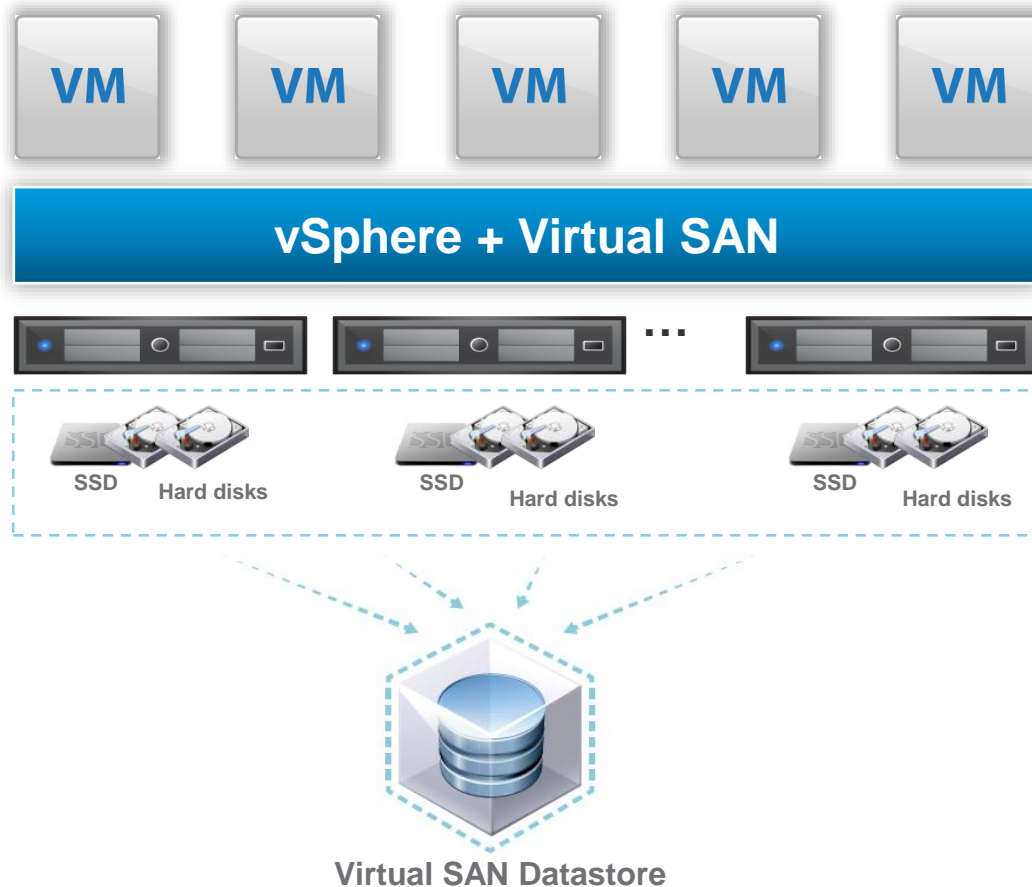


EVO:RAIL



VMware Virtual SAN

Radically Simple Hypervisor-Converged Storage for VMs



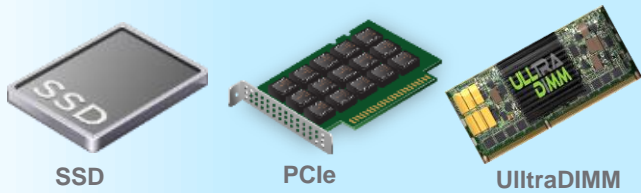
Overview

- Software-defined storage optimized for VMs
- Hypervisor-converged architecture
- Runs on any standard x86 server
- Pools HDD/SSD into a shared datastore
- Delivers enterprise-level scalability and performance
- Managed through per-VM storage policies
- Deeply integrated with the VMware stack

Virtual SAN Can Be Deployed With A Tiered Hybrid Or All-Flash Architecture

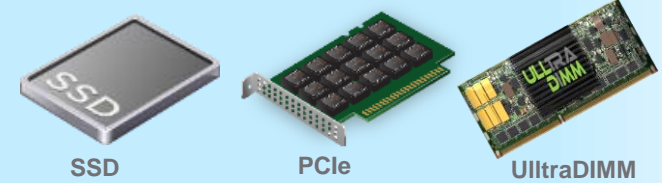


Hybrid



Read and Write Cache

Caching



Writes cached first, Reads go direct to capacity tier

All-Flash



Capacity Tier
SAS/NL SAS/SATA/Direct-attached JBOD

Data Persistence



Capacity Tier
Flash Devices
Reads go directly to capacity tier

40K IOPS per Host*

90K IOPS per Host*
+
sub-millisecond latency

vmware *Performance numbers depend on the workload, randomness, and mix of read/write operation ratios

客戶如何使用VMware Virtual SAN?



Virtual Desktops (VDI)

- Low upfront costs based on commodity x86 servers
- Predictably scale compute and storage with growing user counts



Business Critical Applications

- All-flash, high-performance storage for up to 90K IOPS per host
- Enterprise-class availability with continuous availability



IT Operations

- Deploy management clusters on simple, low TCO infrastructure
- Support IT operations with low-cost, simple storage



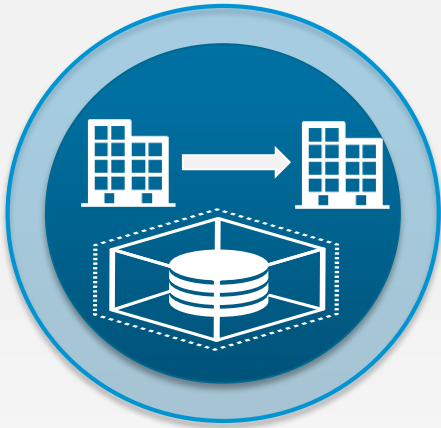
Remote IT (ROBO)

- Powerful, simple storage for limited IT staffs or expertise
- 2-node configuration for low cost, ROBO solution

What is New in 6.1

What's New in Virtual SAN 6.1

Enterprise Availability and Data Protection



- ✓ Stretched Cluster with RPO=0, metro-distance
- ✓ 5 min RPO vSphere Replication
- ✓ Support for SMP-FT
- ✓ Support for Oracle RAC and Microsoft MSCS

Advanced Management & Troubleshooting



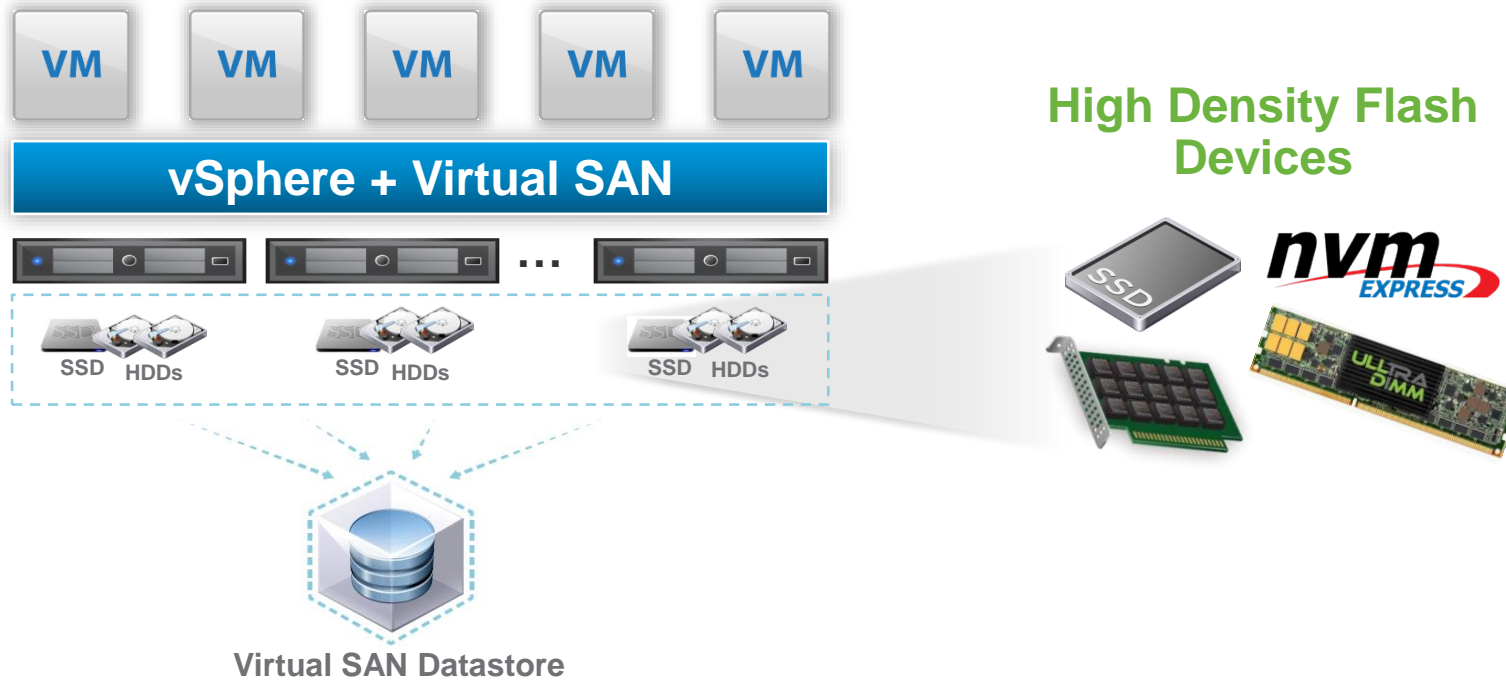
- ✓ Health Check plug-in for HW monitoring, compliance
- ✓ vRealize Operations integration for capacity planning and root-cause analysis
- ✓ Support cloud-native apps

New Hardware Options



- ✓ 2-node clusters for ROBO
- ✓ New Ready Nodes
- ✓ New SSD HW options:
 - Intel NVMe
 - Diablo Ultra DIMM

New Flash Hardware Devices Supported



- Less than $5 \mu\text{s}$ write latency: 3x improvement vs external arrays
- Deploy Virtual SAN in thin blade form factor
- Achieve $\sim 100\text{k}$ IOPs/host with NVMe

- **SanDisk UltraDIMM™** SSDs connect flash storage to the memory channel via DIMM slots, achieving very low ($<5\mu\text{s}$) write latency
- **NVMe** allows for greater parallelism to be utilized by both hardware and software and as a result various performance improvements

World's 1st 64-Node, All-Flash Array with Virtual SAN and NVMe

See VMware Virtual SAN live and at scale in the HCI Zone!

64

Nodes

6,400

VMs

4.2M

IOPS

500

Terabytes



Where:
HCI Zone

When:
Expo Hours

Sponsor:
Intel

設計

使用單位的需求

- 160 vm , each vm need 1 vmdk , each vmdk = 320GB , Number Of Failures To Tolerate = 1 , where can I start?
- The capacity
 - How to calculate ?
 - [1 SSD + 7 HDD (SSD)]/ disk group,
 - Each host maximum 5 disk groups (5 + 35)
 - Boot partition
- The physical host
 - deploying ESXi hosts with similar or identical configurations across all cluster members, including similar or identical storage configurations
 - While hosts that do not contribute storage can still leverage the Virtual SAN datastore if they are part of the same vSphere cluster , VMware is not recommending unbalanced configurations.
 - 4 node clusters allow for greater flexibility. Consider designing clusters with a minimum of 4 nodes where possible.

設計與評估的主要思考方向



Hardware Selection & Design

Always use **certified Ready Nodes**

Pick Ready Node series based on expected IOPS, VM density & capacity

Use a **balanced** configuration

Always use **latest VMware HCL certified versions of firmware & driver** for controllers

Design for availability & future growth

Size for Performance & Capacity

Ensure **SSD:HDD ratio** is **1:10** of usable capacity

Follow Ready Node guidance to pick right class of drives (HDD & SSD)

Pick 8-Series RNs for performance intensive workloads

Use **SAS Expander** based Ready Nodes for **capacity intensive** workloads. SAS expanders are certified on a per platform basis. **Ensure your Ready Node is certified for Expanders**

Recommended Best Practices

SAS or NL-SAS recommended over SATA for performance & reliability

Avoid vmfs data stores on boot devices which are behind the same controller as VSAN data store

Ensure controller **queue depth > 256** for performance & stability. **Pass through** recommended over RAID 0

10G Network recommended for any Ready Node above 2-Series. **Multicast required** across all hosts. Enable **jumbo frames/NIC teaming** for higher performance/redundancy

概念驗證

概念驗證重點

- What are the most important test validation?
 1. Successful VSAN configuration
 2. Successful VM deployments on VSAN datastore
 3. VM Availability in the event of failures (host, storage device, network)
 4. VSAN serviceability
 5. VM Performance meets expectations

Advanced Troubleshooting with Virtual SAN Health Check Plug-in

Free tool designed to deliver troubleshooting and health reports about Virtual SAN subsystems

The screenshot shows the VMware vSphere Web Client interface. The left-hand side contains a 'Navigator' pane with a tree view of hosts and clusters. The main content area is titled 'MGMT' and shows the 'Monitor' tab. Under the 'Monitor' tab, there are several sub-tabs: 'Issues', 'Performance', 'Profile Compliance', 'Health', 'Tasks', 'Events', 'Resource Reservation', 'Virtual SAN', 'vSphere DRS', and 'Utilization'. The 'Virtual SAN' sub-tab is selected, and the 'Vsan Health Tests' section is displayed. The 'Vsan Health Tests' section shows a table of test results with columns for 'Test Name' and 'Status'. The overall health is reported as 'OK'.

Test Name	Status
VSAN Health Service update-to-date	✓ OK
Advanced Virtual SAN configuration in sync	✓ OK
Limits health	✓ OK
Current cluster situation	✓ OK
After 1 additional host failure	✓ OK
Virtual SAN object health	✓ OK
Network health	✓ OK
Hosts disconnected from VC	✓ OK
Hosts with connectivity issues	✓ OK
VSAN cluster partition	✓ OK
Unexpected VSAN cluster members	✓ OK
VSAN cluster partition	✓ OK
Hosts with VSAN disabled	✓ OK
All hosts have a VSAN vmknic configured	✓ OK
All hosts have matching subnets	✓ OK
All hosts have matching multicast settings	✓ OK
Hosts small ping test (connectivity check)	✓ OK
Hosts large ping test (MTU check)	✓ OK
Multicast assessment based on other checks	✓ OK
Physical VSAN disk related health	✓ OK
Physical VSAN disks	✓ OK
Component metadata health	✓ OK
Memory pools (heaps)	✓ OK
Memory pools (slabs)	✓ OK

- Cluster Health
- Network Health
- Data Health
- Limits Health
- Physical Disk Health

項目,敘述,驗證步驟,預期結果,測試結果

ITEM no.	Function	Description	validation step	result	Notice
21	Object Failures	RAID1 - Secondary VM component failure	<ol style="list-style-type: none"> 1. Start a 4 host vsan cluster. 2. When configuring all of the hosts' networks, add one network for VMotion, one network for FT Logging and one vsan network. 3. Provision a Windows/Linux VM on vsan datastore with RAID1 configuration (HFFT =1, SW =1). 4. Turn on FT on the powered on VM and verify the vsan components and storage policy information on both primary and secondary VMs. 5. Inject permanent disk errors on one of secondary VM vdisk's vsan component owner disk. 6. Perform FT VM failover. 7. VM should successfully failed over. 8. New secondary is spawned on another host. 9. Clear the disk errors and the failed components should be resynced successfully. 		
22		RAID1 - Primary VM component failure	<ol style="list-style-type: none"> 1. Start a 4 host vsan cluster. 2. When configuring all of the hosts' networks, add one network for VMotion, one network for FT Logging and one vsan network. 3. Provision a Windows/Linux VM on vsan datastore with RAID1 configuration (HFFT =1, SW =1). 4. Turn on FT on the powered on VM and verify the vsan components and storage policy information on both primary and secondary VMs. 5. Inject permanent disk errors on one of primary VM vdisk's vsan component owner disk. 6. VM guest should be running without any failures. 7. Perform FT VM failover. 8. VM should successfully failed and the VM guest should be up and running. 9. New secondary is spawned on another host. 10. Clear the disk errors and the failed components should be resynced successfully. 		

佈署與監控

Virtual SAN Performance Troubleshooting

- Virtual SAN Observer is adequate in most cases
- esxtop / VCOps useful in some situations when Observer is not available
 - VM/VCPU bottlenecks
 - Virtual Disk latency vs. Physical Disk latency is a good starting point
- Note: Virtual SAN is a distributed system
 - Performance issues on one host might actually originate on another host
 - Monitor ALL nodes in cluster

Virtual SAN Performance--通常問題來自..

- CPU Bottlenecks
- Network issues
- Hotspots in cluster
- Application requirements vs. physical storage capabilities
- VSAN Overhead (e.g., Re-sync operations, Flush timer, Snapshots)

CPU Bottlenecks

- *症狀:*
 - VSAN threads utilization close to 100% OR high ready time
 - High CPU utilization/ready times manifesting as latency increases
- *原因:*
 - Ready time from system saturation is common case
 - Very few active virtual disks stressing small number of kernel threads
- *解決方法:*
 - Check CPU overcommitment
 - Try distributing workload across multiple virtual disks to increase parallelism in the kernel

Network Issues

- *症狀:*
 - Increase in latency from the DOM Component Manager Layer (VSAN Disks View) to the DOM Owner layer
 - Non-zero network error counts
- *原因:*
 - At very high IOPS and throughput, network kernel threads can be CPU bottlenecked
 - Network misconfiguration or hardware errors
- *解決方法:*
 - Increasing MTU size to 9000 helps in reducing CPU utilization
 - Trace source of errors and fix them

Hot Spots in the Cluster

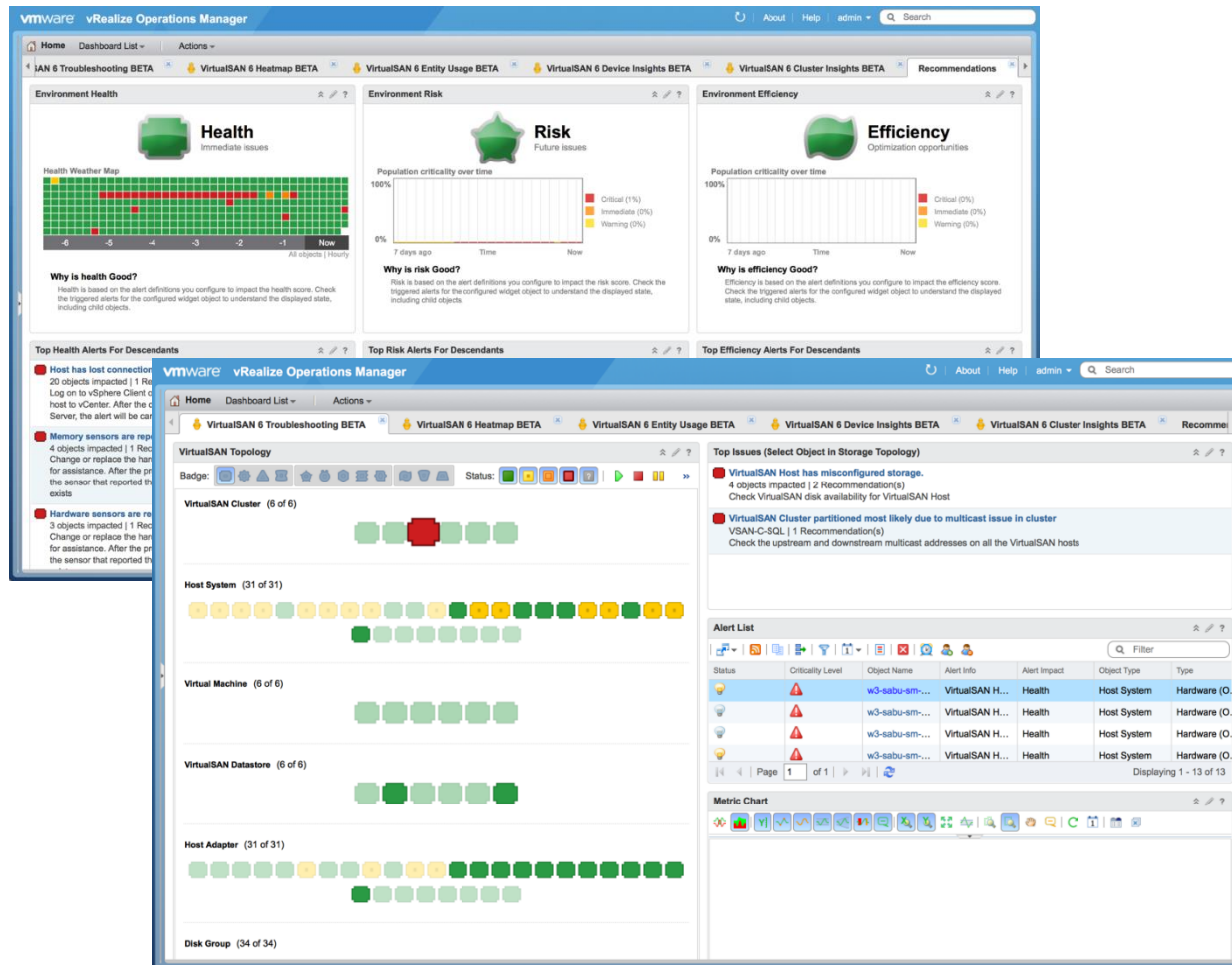
- *症狀:*
 - Subset of hosts (VSAN Disks view) or physical disks (VSAN Disks Deep-Dive) get most of the IOPS
 - Higher load on these hosts/disks → Higher latencies
 - Non-zero congestion levels
- *原因:*
 - Load imbalance in the cluster
 - Virtual disk placement depends on free capacity of physical disks
 - Many disks are provisioned and only few of them are heavily used
- *解決方法:*
 - Set stripeWidth to a value greater than 1 for the heavily utilized disks
 - Proactive disk rebalancing can alleviate the issue

Workload Requirements vs. Physical Disk Capabilities

- *症狀:*
 - High-latency/low-IOPS in virtual disk layer, but no network/CPU issues
 - Non-zero congestion values
 - Low RC Hit rate in Hybrid clusters (VSAN Disks Deep Dive view)
- *原因:*
 - Latency overhead from VSAN, virtualization is more visible at low OIO
 - Read cache, VSAN sparse metadata cache may not be effective for a workload
- *解決方法:*
 - Select hardware based on IOPS requirements
 - Tune cache sizes

Advanced Monitoring, Planning and Troubleshooting with vRealize Operations

Same Dashboards for Easy Virtual SAN Monitoring



- Comprehensive global view across multiple Virtual SAN cluster
- Hundred of KPIs simplified to an easy to consume dashboard
- Smart alerts deliver insight and information – correlate symptoms across the stack

VSAN and vRealize Operations: 降低故障排除的時間

VirtualSAN Cluster partitioned most likely due to multicast issue in cluster

VirtualSAN Cluster partitioned most likely due to non-matching upstream and downstream multicast addresses on all hosts in the cluster

Recommendations

Check the upstream and downstream multicast addresses on all the VirtualSAN hosts

Alert Information

Object Name: VSAN-C-SQL
Control State: Open
Assigned User: -
Alert Type: Hardware (OSI)
Alert Subtype: Availability
Status: Active
Impact: Health
Criticality: Critical
Start Time: 4/28/15 7:10 PM
Update Time: 4/28/15 7:10 PM
Cancel Time:

What is Causing the Issue ?

VirtualSanCluster VSAN-C-SQ...

VirtualSanCluster VSAN-C-SQ...

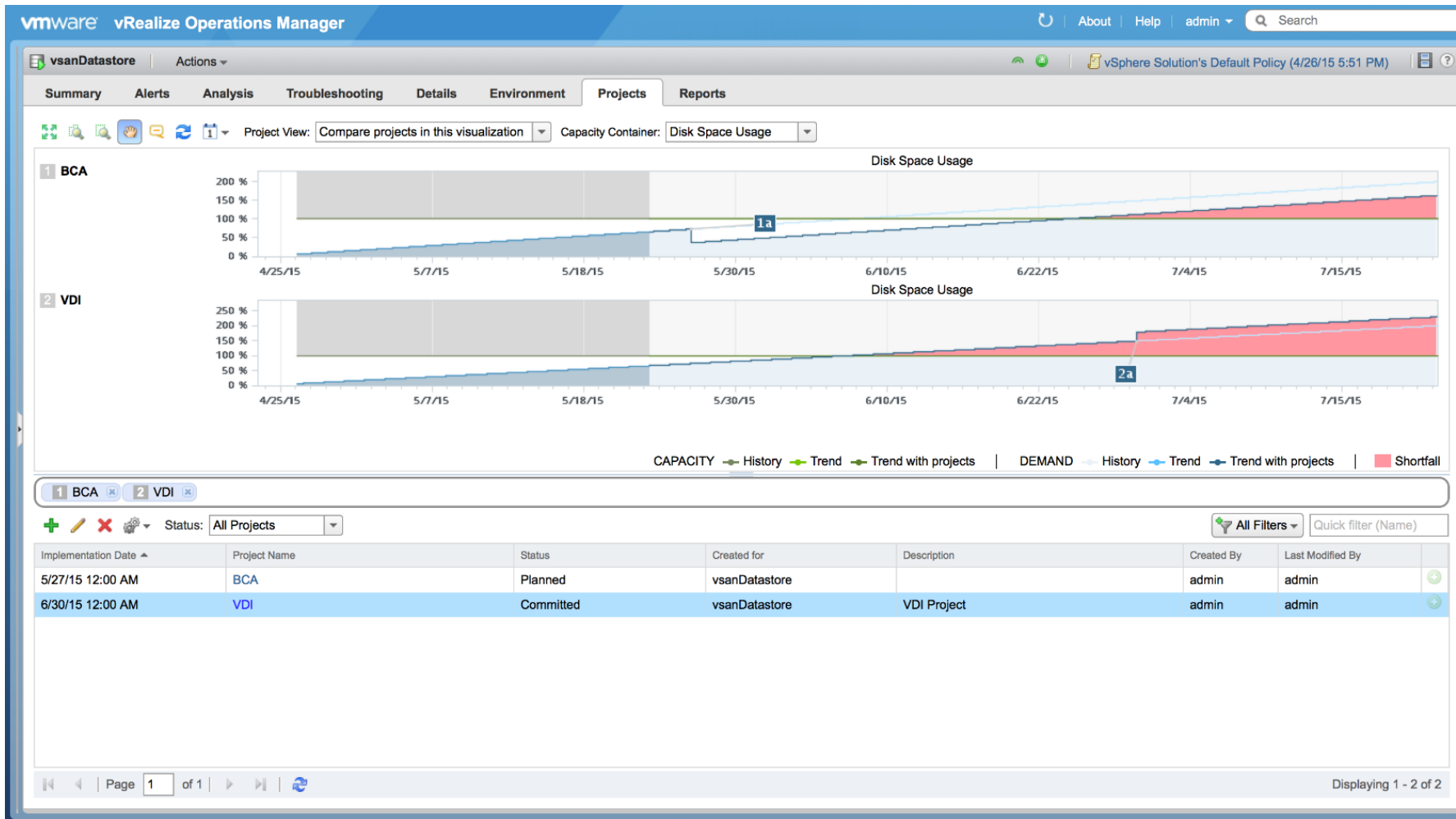
Event source: VSAN-C-SQL
Source event object name: VirtualSanCluster VSAN-C-SQL detected network partitioning
Source event name: VirtualSanCluster VSAN-C-SQL detected network partitioning
Source event status: All hosts in the cluster unable to communicate due to partitioning.
Device Description: [w3-sabu-sm-010.eng.vmware.com] are partitioned [w3-sabu-sm-009.eng.vmware.com] are partitioned [w3-sabu-sm-011.eng.vmware.com] are partitioned [w3-sabu-sm-012.eng.vmware.com] are partitioned

Event source: VSAN-C-SQL
Source event object name: VirtualSanCluster VSAN-C-SQL detected multicast address issue
Source event name: VirtualSanCluster VSAN-C-SQL detected multicast address issue
Source event status: Hosts in the cluster have different Multicast addresses set

- Get prescriptive guidance for remediation, including automated actions
- Multiple symptoms combined to provide an in-depth root cause analysis

VSAN and vRealize Operations: Capacity Planning

Never run out of or overprovision capacity again



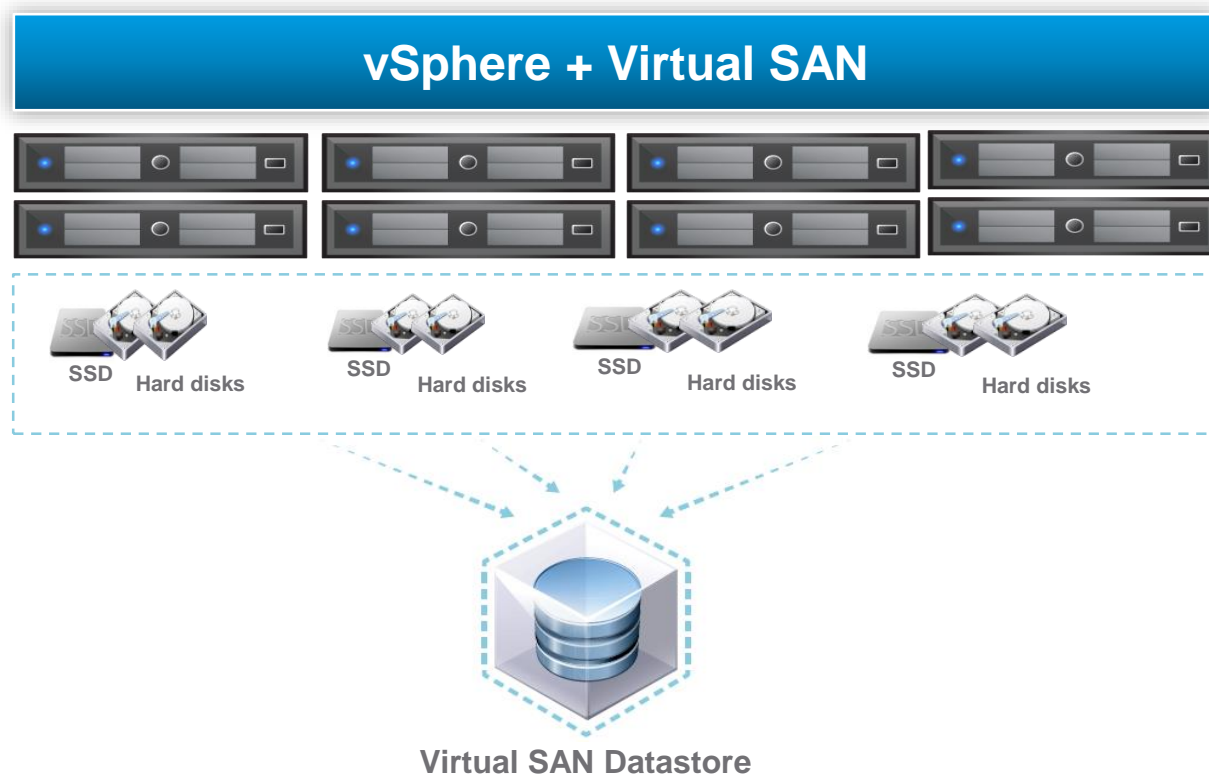
- Plan capacity consumption based on current consumption and future projects
- Monitor capacity usage and identify overprovisioned resources
- Enhanced “what-if” scenarios and alert settings

案例分享



案例

Virtualizing Microsoft Applications on VMware Virtual SAN



Takeaways

- 8 node Hybrid VMware Virtual SAN Cluster
- 4 Exchange 2013 Mailbox Servers - DAG
- 4 Exchange 2013 CAS
- 2 SQL Server 2014 - AAG
- SharePoint 2013
- Windows Files Share Hosted on VMware Virtual SAN
- No need for Zoning or specialized tools

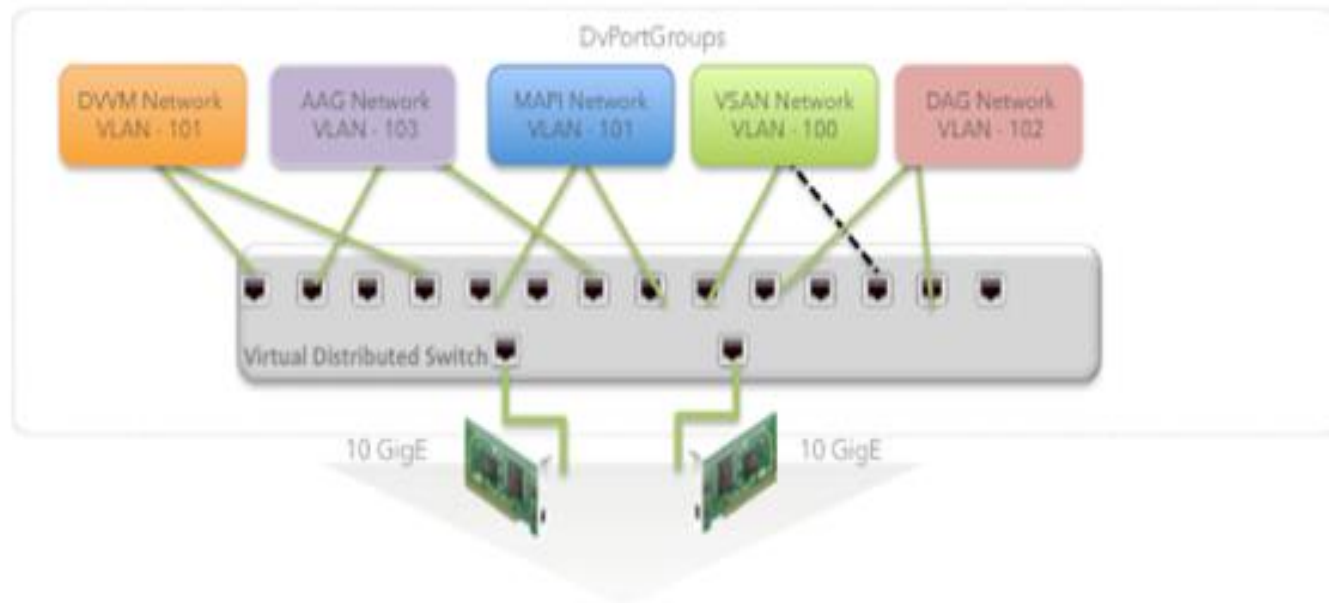
VMware Virtual SAN 實體機規格

COMPONENT	SPECIFICATIONS
ESX host CPU	2 x Intel(R) Xeon(R) CPU E5-2690 v2 @ 3.00GHz 10C (60GHz)
ESX host RAM	256GB
ESX Version	ESXi 6.0.build
Network Adapter	2x 10-Gigabit SFI/SFP+
Storage Controller	2x 12Gbps HBA
Power Management	Balanced (set in BIOS)
Disks	SSD: 2x Intel 400GB SSD: 2x Intel 200GB HDD: 12 x Seagate 900GB

Takeaways

- Commodity Hardware
- HCL Supported HBAs
- Solution Sizing is critical

Virtual Networking Configuration



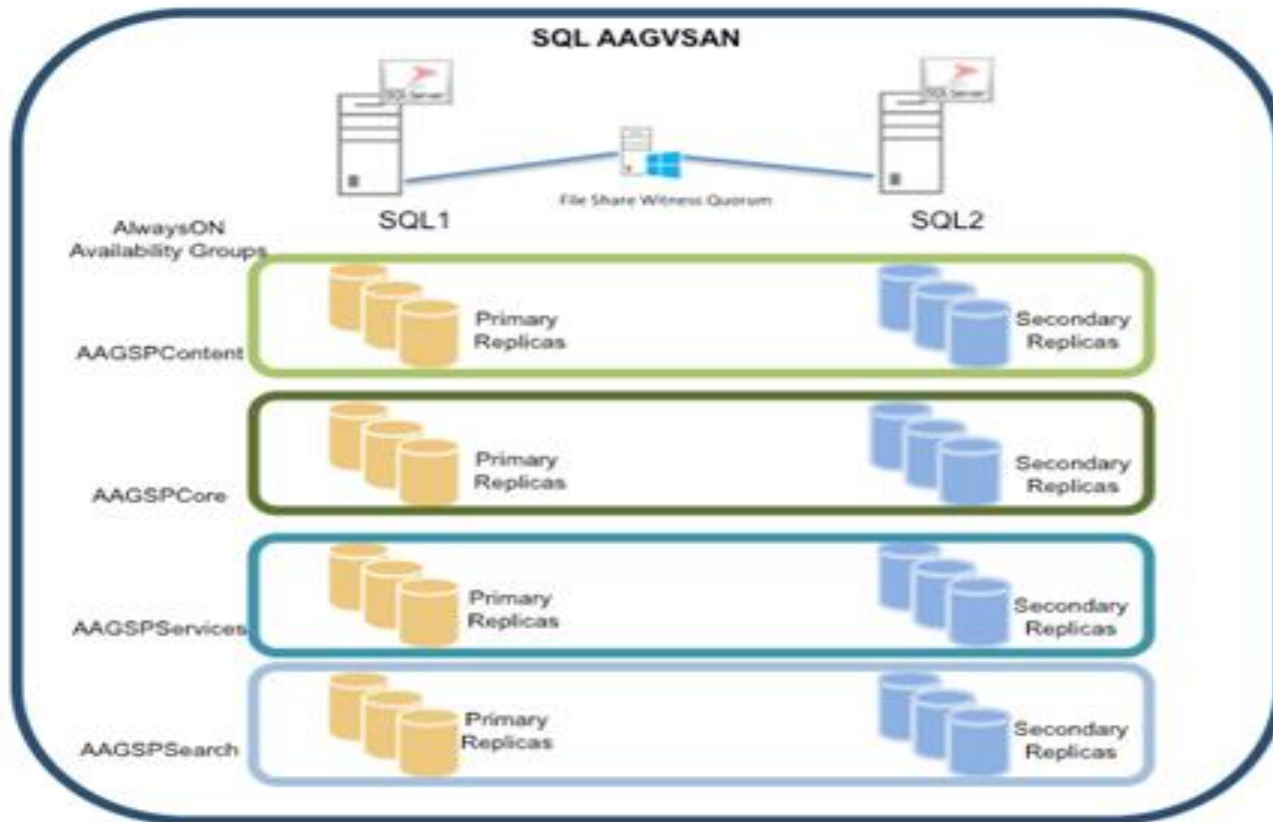
Takeaways

- Refer VMware and Microsoft Best Practices
- Use VMware dVswitch

SQL Configuration & Performance



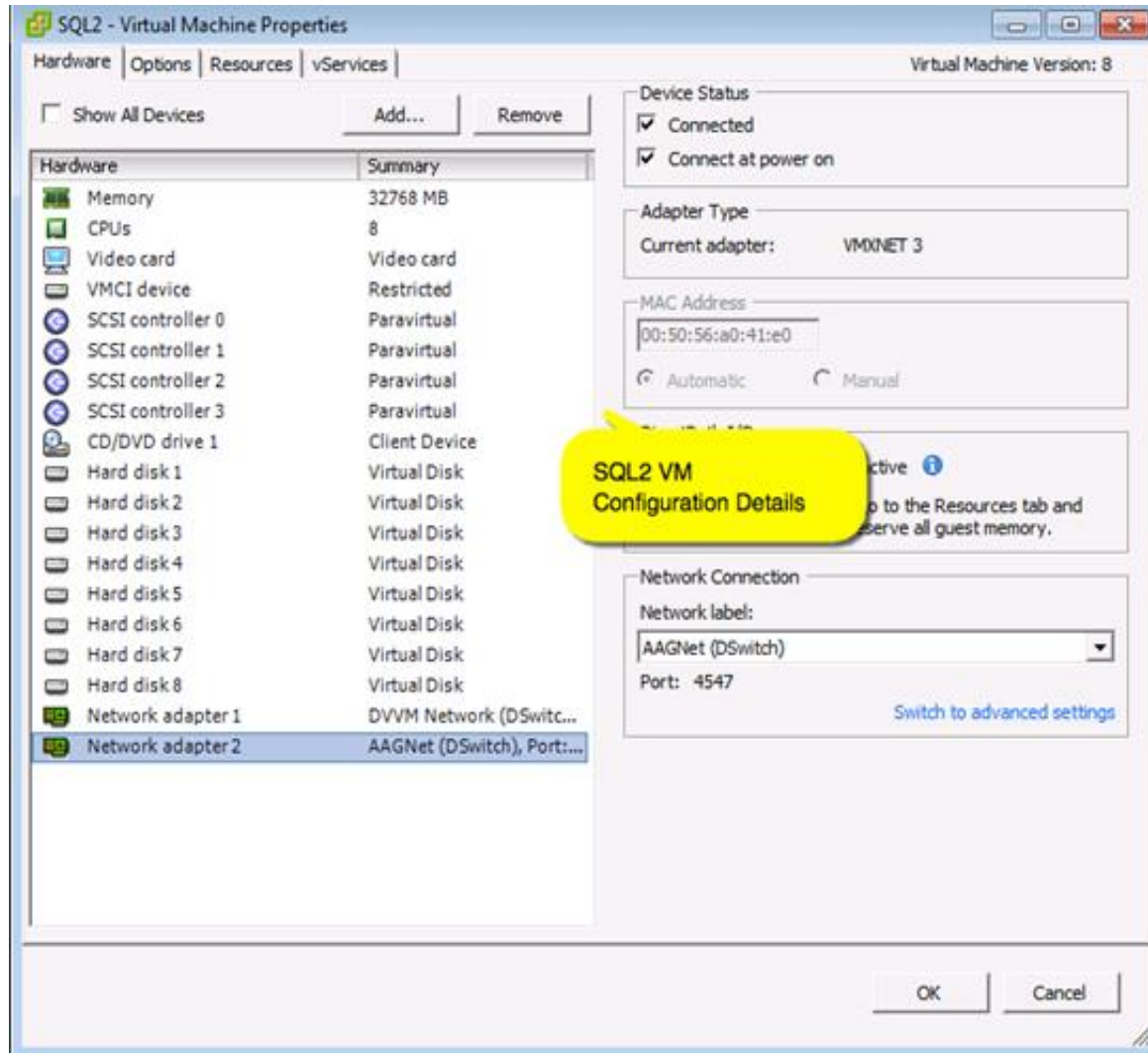
SQL 2014 Architecture



Takeaways

- Follow VMware and Microsoft SQL Server Best Practices
- MSCS SMB Share for AAG

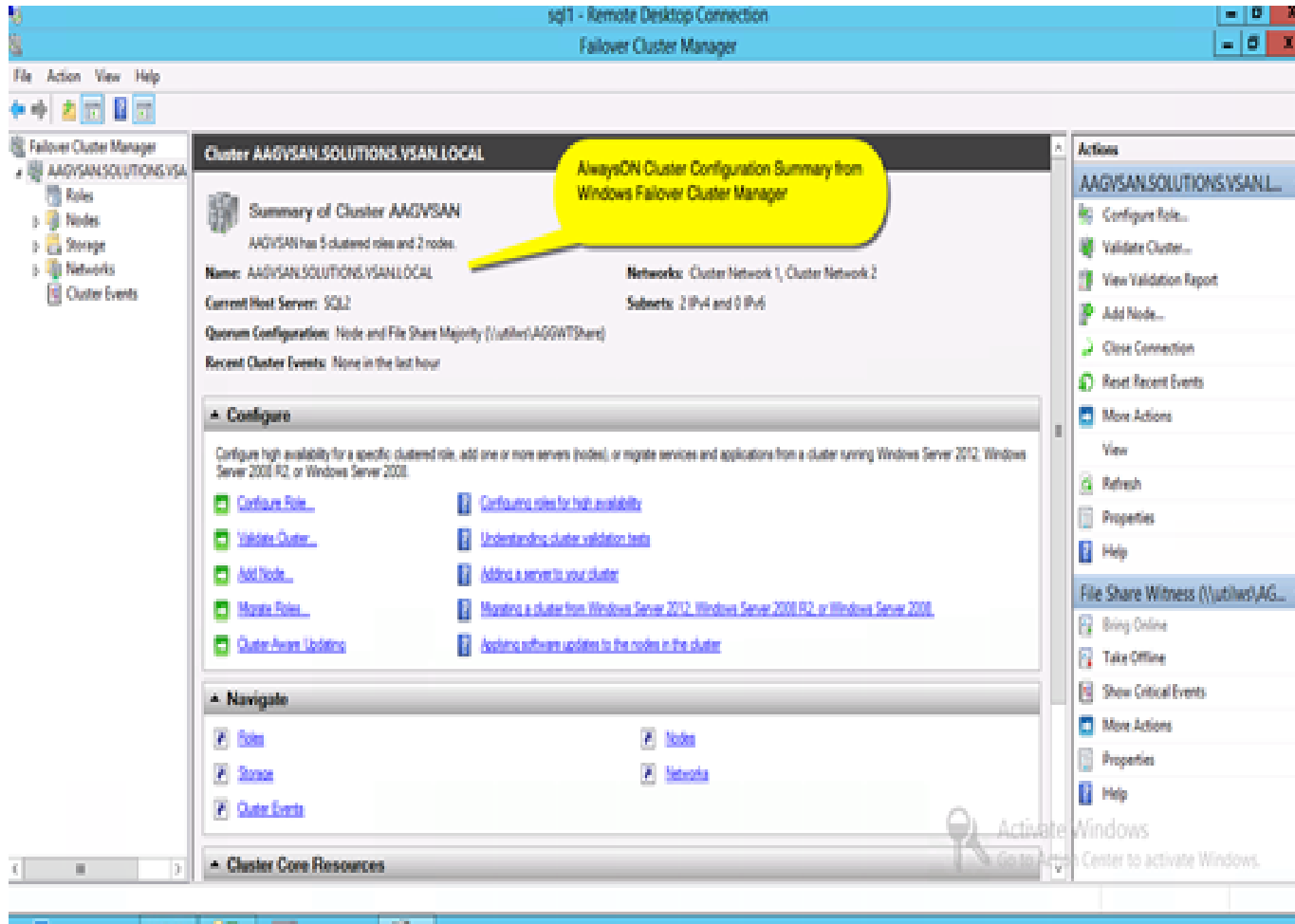
SQL VM Detail



Takeaways

- Refer VMware SQL Server Best Practices
- VM Sizing is critical

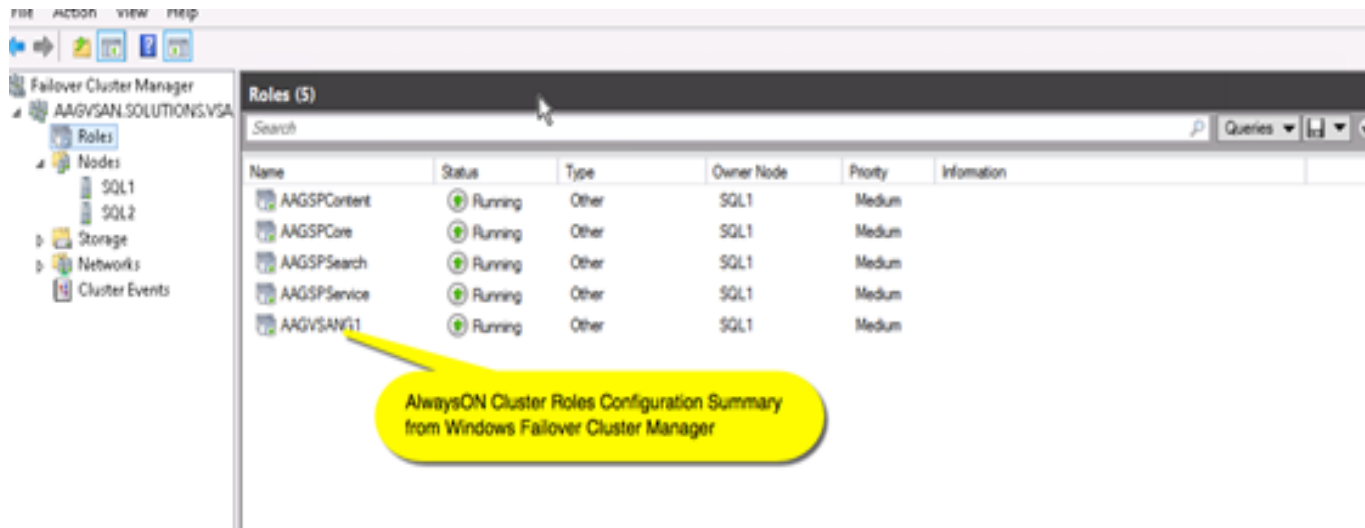
SQL AAG Overview – Windows Failover Cluster Manager



Takeaways

- Always validate your cluster and review warnings.
- Only Associate Disks when needed

SQL AAG Roles – WFCM



The screenshot displays the Windows Failover Cluster Manager interface. The left-hand navigation pane shows the hierarchy: Failover Cluster Manager > AAG/VSAN.SOLUTIONS.VSA > Roles. The main pane, titled 'Roles (5)', contains a table with the following data:

Name	Status	Type	Owner Node	Priority	Information
AAGSPContent	Running	Other	SQL1	Medium	
AAGSPCore	Running	Other	SQL1	Medium	
AAGSPSearch	Running	Other	SQL1	Medium	
AAGSPService	Running	Other	SQL1	Medium	
AAGVSAN01	Running	Other	SQL1	Medium	

A yellow callout bubble points to the 'AAGVSAN01' role with the text: 'AlwaysON Cluster Roles Configuration Summary from Windows Failover Cluster Manager'.

Takeaways

- Review your cluster roles and ensure they reflect what SQL Management Studio reports

SQL AAG – SQL Management Studio

The screenshot shows the SQL Management Studio interface. On the left, the Object Explorer displays a tree view of the server instance. The 'AlwaysOn High Availability' folder is expanded, showing 'Availability Groups'. Under this, there are several primary instances: AAGSPContent (Primary), AAGSPCore (Primary), AAGSPSearch (Primary), AAGSPService (Primary), and AAGVSANG1 (Primary). The right pane shows a dashboard titled 'Availability groups on SQL1'. It contains a table with the following data:

Availability Group Name	Primary Instance	Failover Mode	Issues
AAGSPContent	SQL1	Automatic	
AAGSPCore	SQL1	Automatic	
AAGSPSearch	SQL1	Automatic	
AAGSPService	SQL1	Automatic	
AAGVSANG1	SQL1	Automatic	

A yellow callout bubble points to the table with the text: 'Here is a summary view of the AlwaysON Availability Groups'.

Takeaways

- SQL Management Studio has great detail for AAG...use it

Performance – SQL DVD Store

Combined totals for both test ds2sqlserver sessions:

- Total Purchases during 2 hours: 4969793
- Average Orders Per Minute: 41415

DVD Store Test ds2sqlserver session 1

- Total Purchases during 2 hours: 2484763
- Average Orders Per Minute: 20706

DVD Store Test ds2sqlserver session 2

- Total Purchases during 2 hours: 2485030
- Average Orders Per Minute: 20709

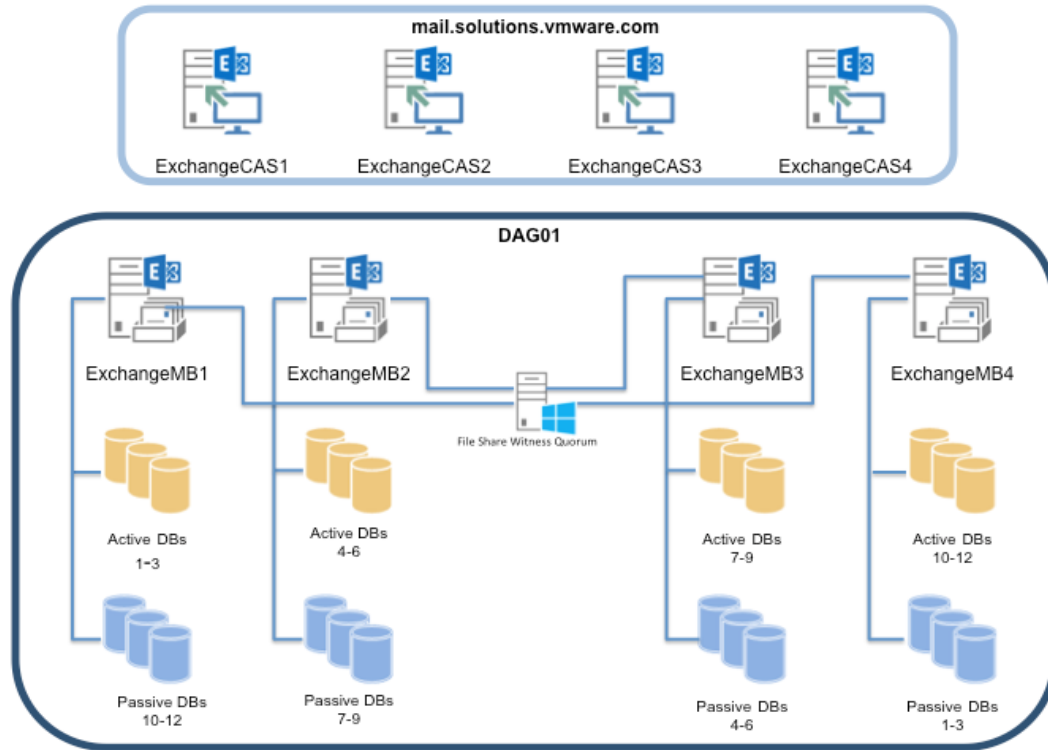
Takeaways

- This is commodity hardware with enterprise performance
- All results are as good if not better than NAS/SAN storage

Exchange Configuration & Performance



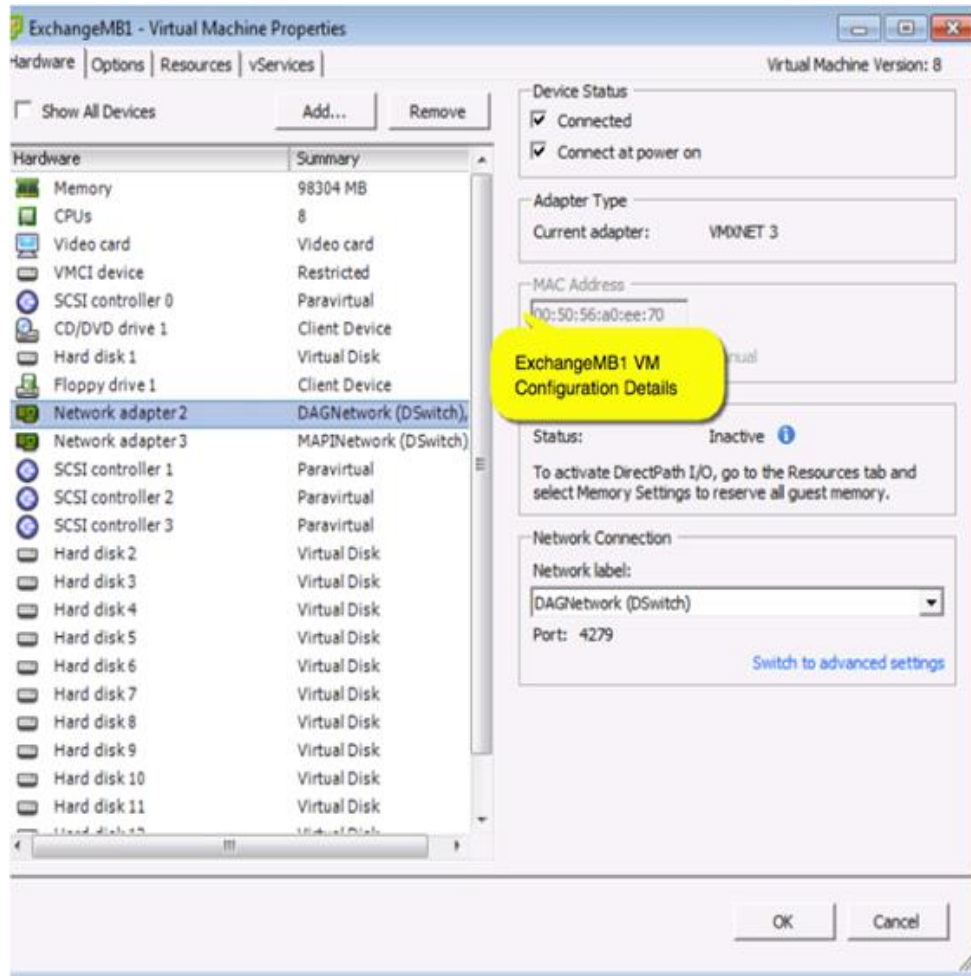
Exchange 2013 Architecture



Takeaways

- Follow VMware and Microsoft Exchange Best Practices
- MSCS SMB Share for DAG

Exchange VM Detail



Takeaways

- Refer VMware Exchange Best Practices
- Use Exchange Sizing Guide
- VM Sizing is critical

Exchange VM Network Detail

Server Manager - Local Server

PROPERTIES For ExchangeMB1

Computer name	ExchangeMB1	Last installed updates	3/16/2015 3:07 PM
Domain	SOLUTIONS.VSAN.LOCAL	Windows Update	Install updates auto
Cluster name	DAG1	Last checked for updates	3/17/2015 10:03 AM
Cluster object type	Cluster Node		
Windows Firewall	Domain: Off, Public: Off	Windows Error Reporting	Off
Remote management	Enabled	Customer Experience Improvement Program	Not participating
Remote Desktop	Enabled	IE Enhanced Security Configuration	On
NIC Teaming	Disabled	Time zone	(UTC-08:00) Pacific
DAG	10.10.10.11	Product ID	00184-90000-00001
MAPI	192.168.100.141		
Operating system version	Microsoft Windows Server 2012 Datacenter	Processors	Intel(R) Xeon(R) CPL
Hardware information	VMware, Inc. VMware Virtual Platform	Installed memory (RAM)	96 GB
		Total disk space	6779.62 GB

Note MAPI and DAG networks

Takeaways

- Keep MAPI and DAG traffic separate
- Name NICs

Exchange VM Network Detail – Continued

ReplicationDagNetwork01 Help

*Database availability group network name:
ReplicationDagNetwork01

Description:

Subnets:
+ ✎ -

SUBNET	STATUS
10.0.0.0/8	Up

Network interfaces:

NETWORK INTERFACE	STATUS
10.10.10.11	Up
10.10.10.12	Up
10.10.10.13	Up
10.10.10.14	Up

Enable replication

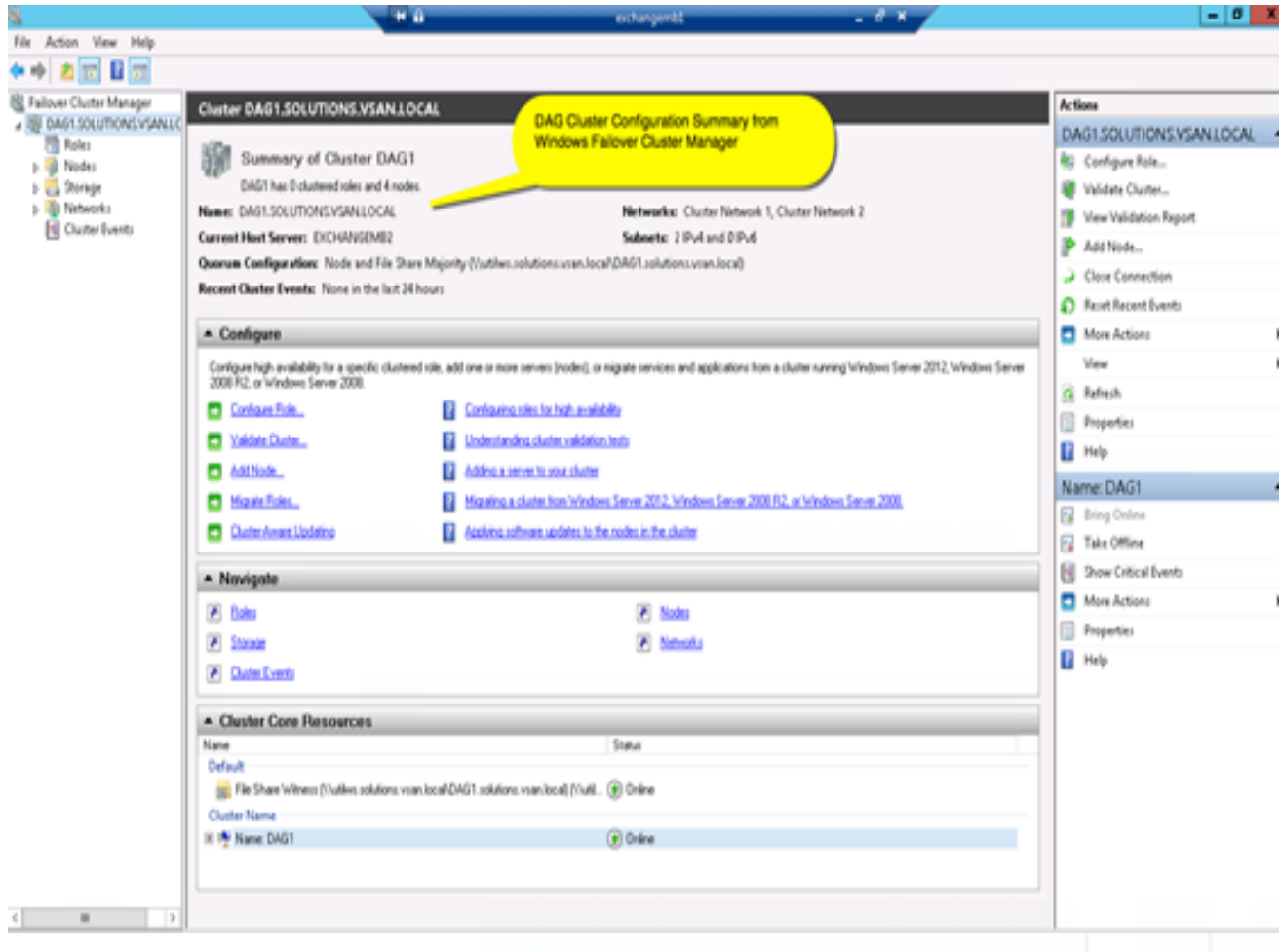
Use this field to specify a name for the DAG network of up to 128 characters. The name of the network must be unique within the DAG.

DAG Network Detail from Exchange Control Panel

Takeaways

- Configure Exchange DAG as per Best Practices

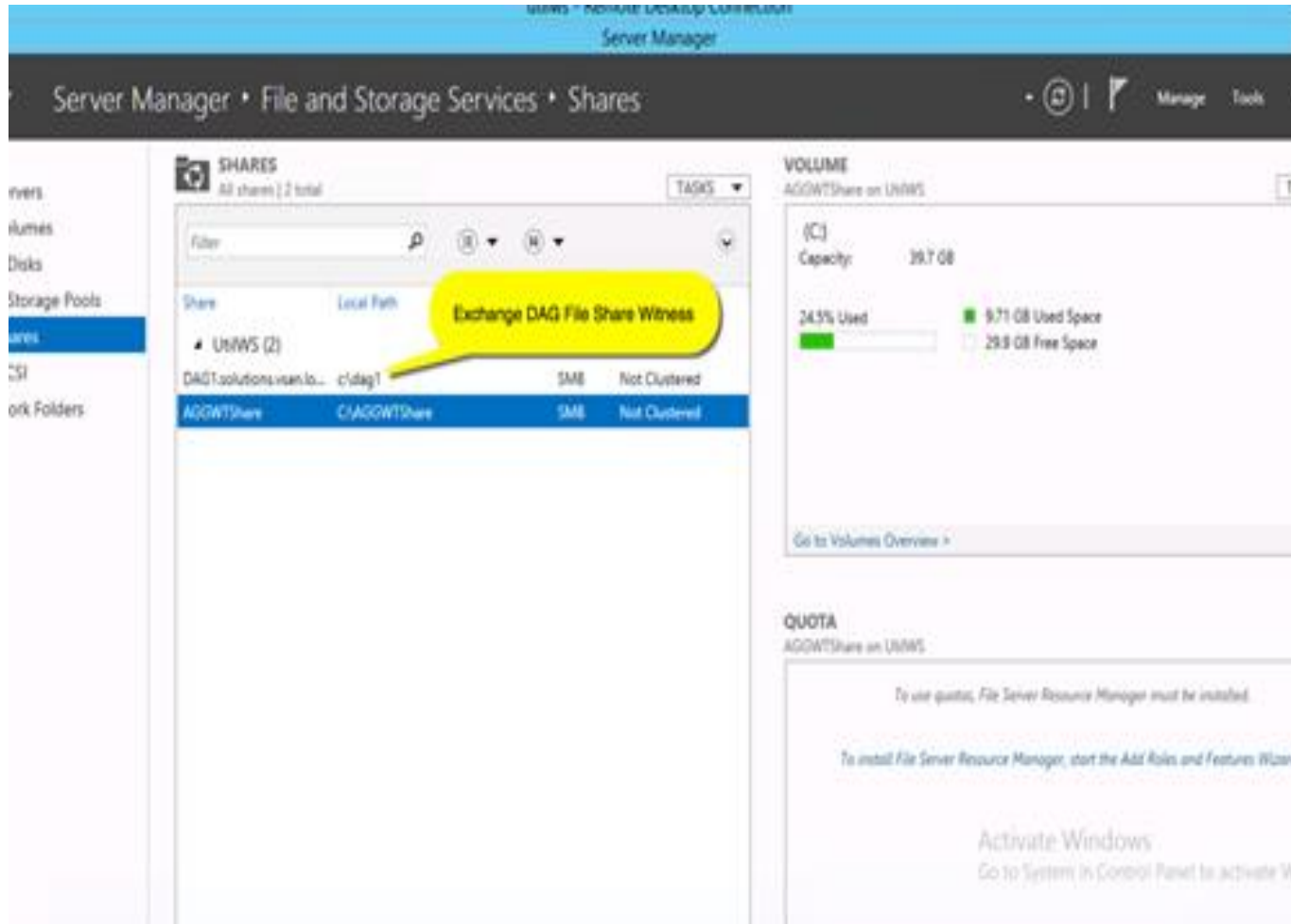
Exchange DAG Overview – Windows Failover Cluster Manager



Takeaways

- Always validate your cluster and review warnings.
- Only Associate Disks when needed

Windows File Share Witness



Takeaways

- Use Standalone Windows Server VM as SMB Share

Performance – Exchange Jetstress Server MB1

PERFORMANCE COUNTERS	TARGET VALUES
Achieved Exchange transactional IOPS (I/O database reads/sec + I/O database writes/sec)	1194
I/O database reads/sec	132
I/O database writes/sec	66
Total IOPS (I/O database reads/sec + I/O database writes/sec + BDM reads/sec + I/O log replication reads/sec + I/O log writes/sec)	325
I/O database reads average latency (ms)	Less than 20 ms
I/O log reads average latency (ms)	Less than 10 ms

Takeaways

- Results are on well below the max latencies

Performance – Exchange Jetstress Server MB2

PERFORMANCE COUNTERS	TARGET VALUES
Achieved Exchange transactional IOPS (I/O database reads/sec + I/O database writes/sec)	1265
I/O database reads/sec	140
I/O database writes/sec	70
Total IOPS (I/O database reads/sec + I/O database writes/sec + BDM reads/sec + I/O log replication reads/sec + I/O log writes/sec)	325
I/O database reads average latency (ms)	Less than 20 ms
I/O log reads average latency (ms)	Less than 10 ms

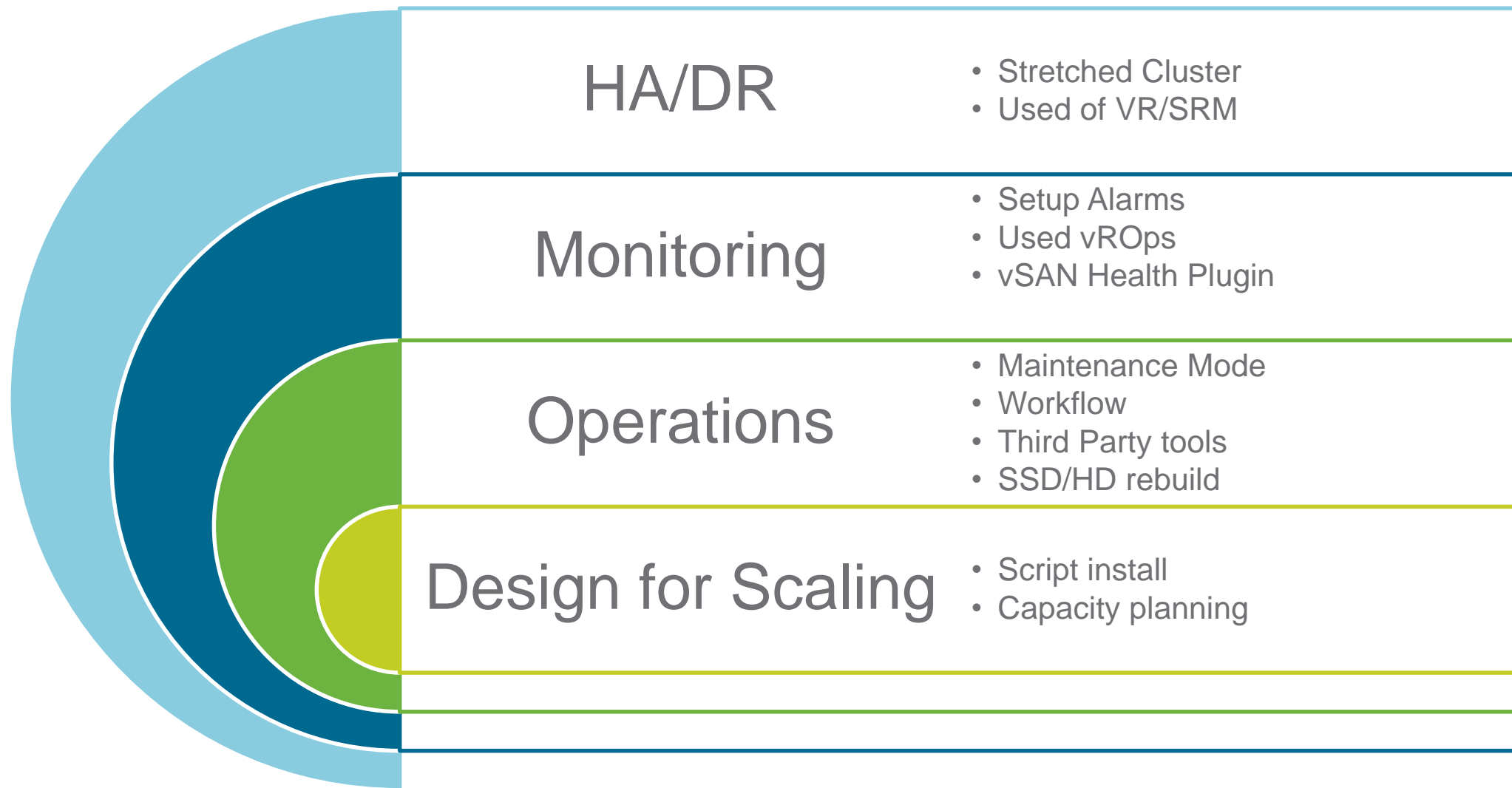
Takeaways

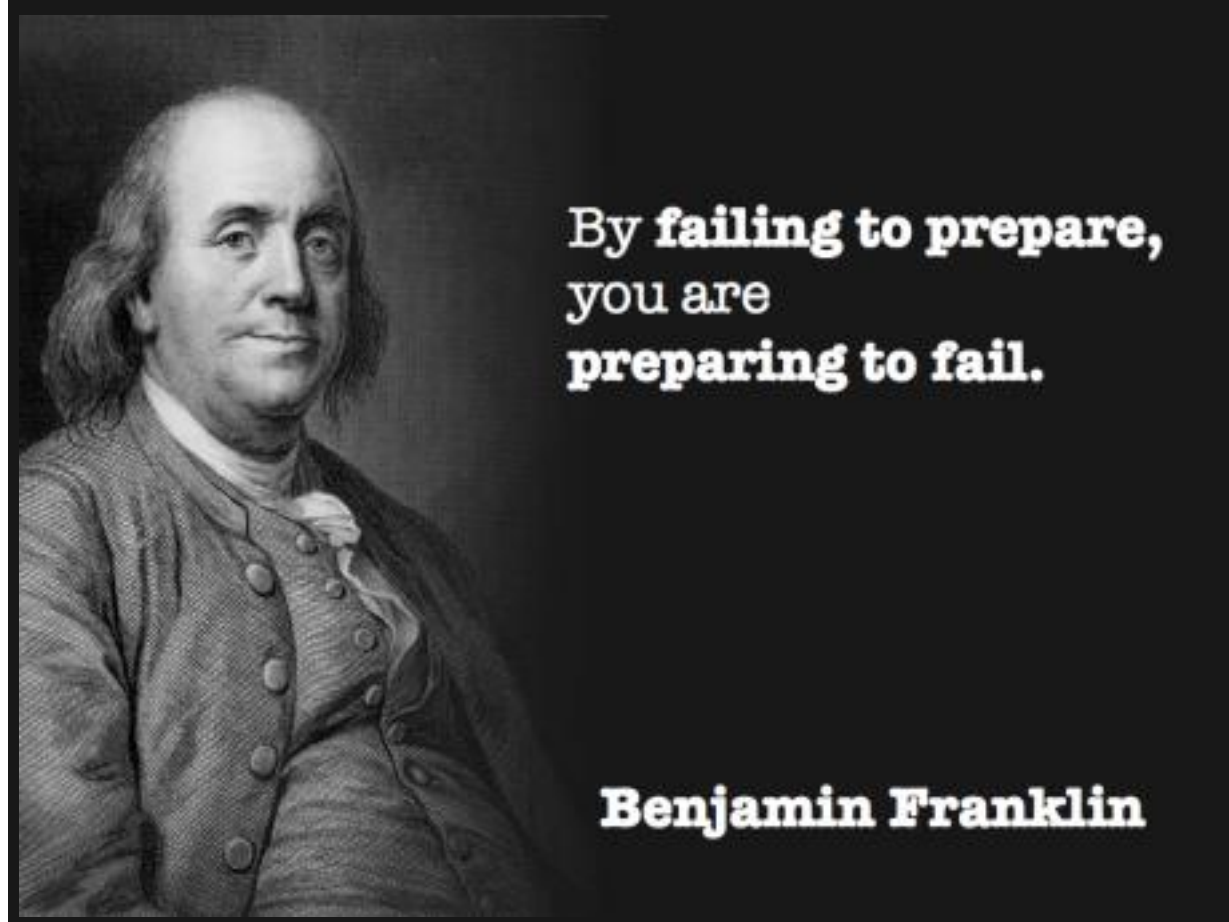
- This is commodity hardware with enterprise performance
- All results are as good if not better than NAS/SAN storage

總結



注意事項





**By failing to prepare,
you are
preparing to fail.**

Benjamin Franklin

Links

- Download: https://my.vmware.com/group/vmware/info/slug/datacenter_cloud_infrastructure/vmware_virtual_san/6_0
- Doc landing page: <http://www.vmware.com/support/pubs/virtual-san-pubs.html>
- <https://www.youtube.com/playlist?list=PL9MeVsU0uG65kM9iszb5KmNj01PiAWgvf>
- Admin guide: <http://pubs.vmware.com/vsphere-60/topic/com.vmware.vsphere.virtualsan.doc/GUID-AEF15062-1ED9-4E2B-BA12-A5CE0932B976.html>
- Product page: <http://www.vmware.com/products/virtual-san/>
- TCO Calculator: <https://vsantco.vmware.com/vsan/SI/SIEV>
- VSAN ReadyNode: <http://partnerweb.vmware.com/programs/vsan/Virtual%20SAN%20Ready%20Nodes.pdf>
- <http://blogs.vmware.com/vsphere/storage>
- <http://www.yellow-bricks.com/virtual-san/>
- perf: <http://www.vmware.com/files/pdf/products/vsan/VMware-Virtual-San6-Scalability-Performance-Paper.pdf>
- Design and Sizing: http://www.vmware.com/files/pdf/products/vsan/VSAN_Design_and_Sizing_Guide.pdf
- Troubleshooting: <http://www.vmware.com/files/pdf/products/vsan/VSAN-Troubleshooting-Reference-Manual.pdf>

READY
FOR **ANY**
vForum2015