

# Disponibilidad Y Federación En El Nuevo Centro de Datos



**VMUG**  
VMWARE USER GROUP



Barcelona, 20/01/2012

Ignacio Borrero vSpecialist EMEA Technical  
[ignacio.borrero@emc.com](mailto:ignacio.borrero@emc.com)

# A couple of things to set the stage...

- EMC and VMware – seeing lots of confusion out there re: Disaster Recovery (DR) and Disaster Avoidance (DA)
- Will break this session into multiple parts:
  - PART I – Understanding DR and DA
  - PART II – Understanding Stretched vSphere Clusters
  - PART III – What's New?
  - PART IV – Where are areas where we are working for the future?
- Will work hard to cover a lot, but leave time for Q&A

# PART I...

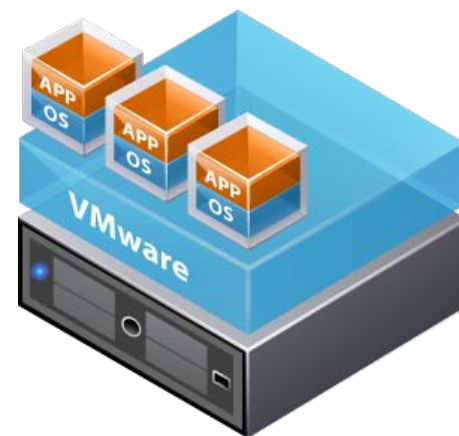
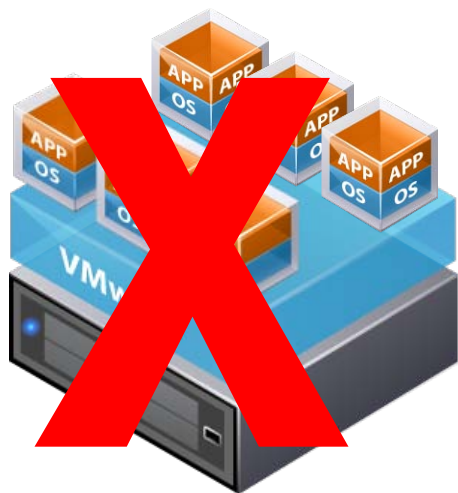
Understanding DR and DA

# “Disaster” Avoidance – Host Level

This is vMotion.

Most important characteristics:

- By definition, avoidance, not recovery.
- *“non-disruptive”* is massively different than *“almost non-disruptive”*



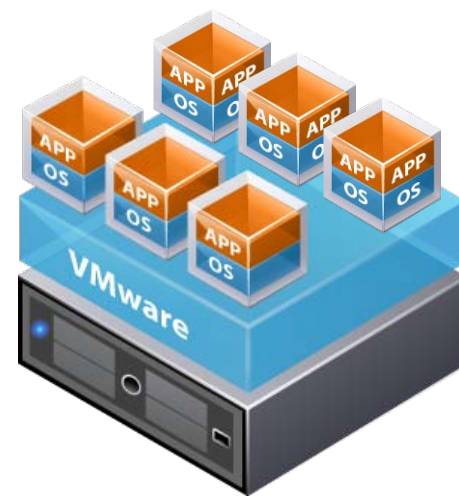
“Hey... That host **WILL** need to go down for maintenance. Let’s vMotion to avoid a disaster and outage.”

# “Disaster” Recovery – Host Level

This is VM HA.

Most important characteristics:

- By definition recovery (restart), not avoidance
- *Simplicity, automation, sequencing*



Hey... That host **WENT** down due to unplanned failure causing a unplanned outage due to that disaster. Let's automate the **RESTART** of the affected VMs on another host.

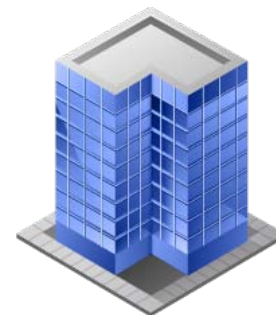
# Disaster Avoidance – Site Level



This is inter-site  
vMotion.

Most important characteristics:

- By definition, avoidance, not recovery.
- *“non-disruptive”* is massively different than *“almost non-disruptive”*



Hey... That site **WILL** need to go down for maintenance. Let's vMotion to avoid a disaster and outage.

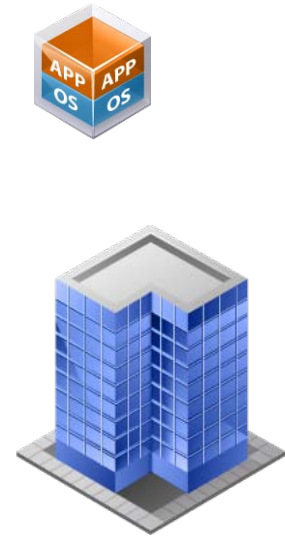
# Disaster Recovery – Site Level



This is Disaster Recovery.

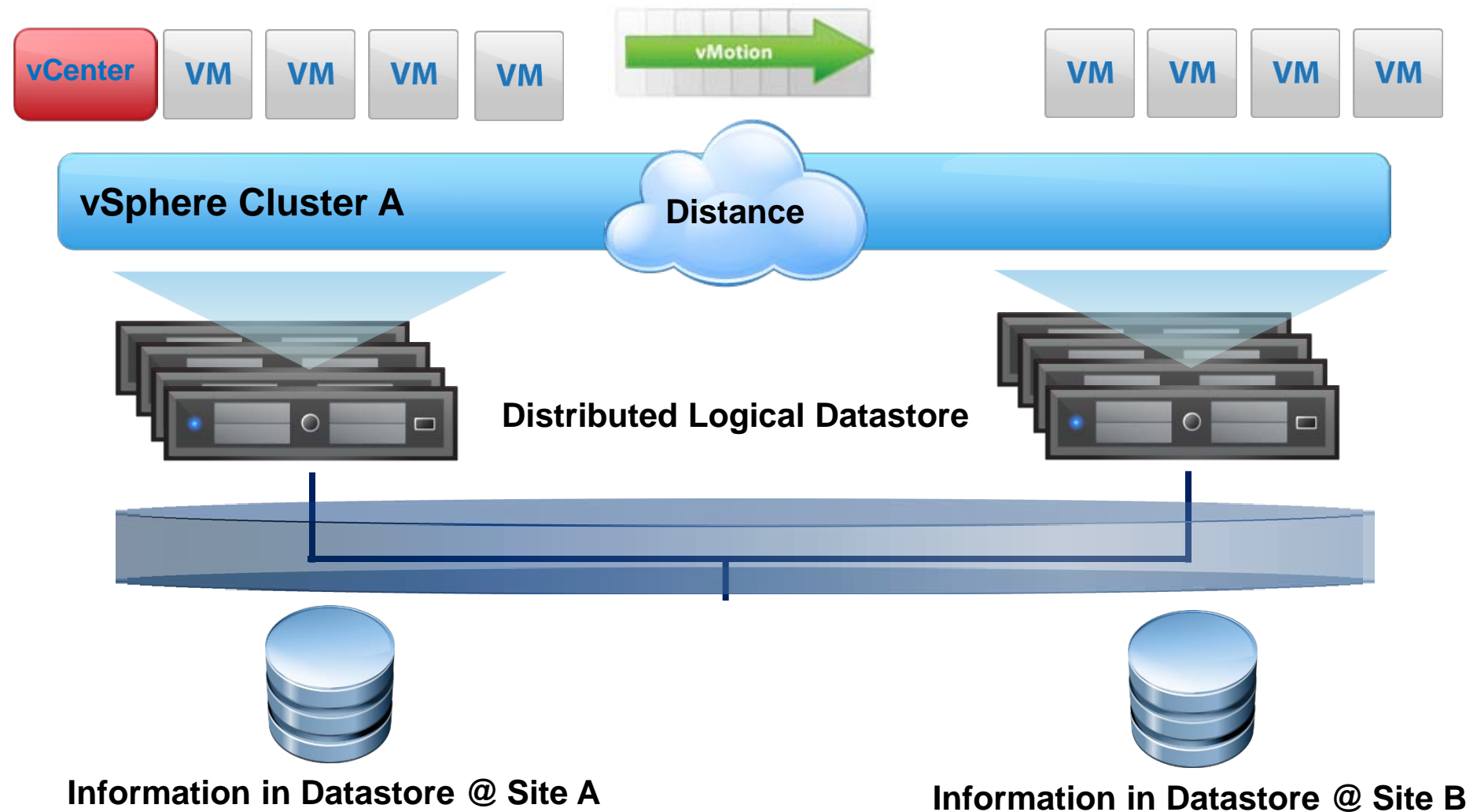
Most important characteristics:

- By definition recovery (restart), not avoidance
- *Simplicity, testing, split brain behavior, automation, sequencing, IP address changes*



Hey... That site **WENT** down due to unplanned failure causing an unplanned outage due to that disaster. Let's automate the **RESTART** of the affected VMs on another host.

# Type 1: “Stretched Single vSphere Cluster”

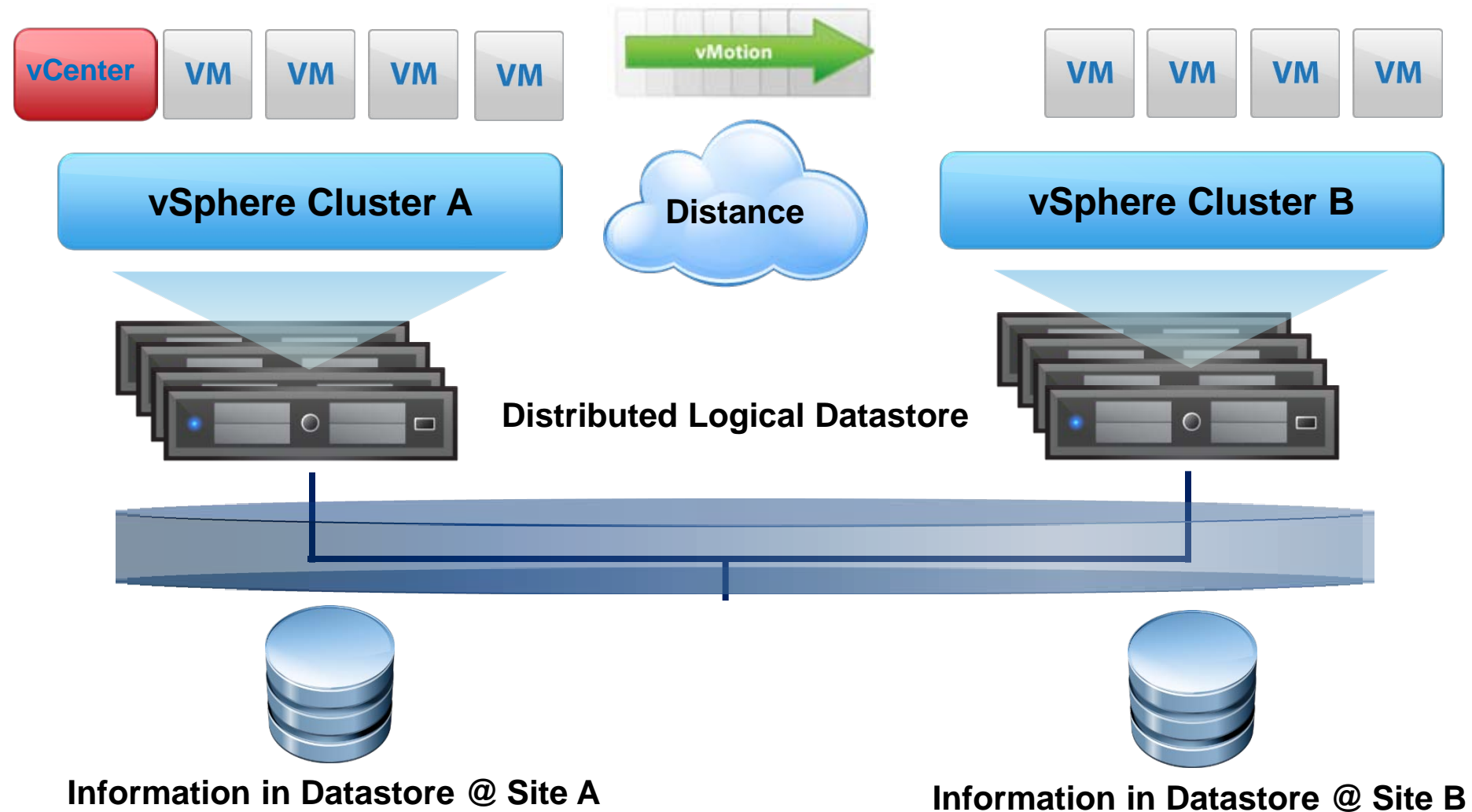




# One little note re: “Intra-Cluster” vMotion

- **Intra-cluster vMotions can be highly parallelized**
  - and more and more with each passing vSphere release
  - With vSphere 4.1 and vSphere 5 it's up to 4 per host/128 per datastore if using 1GbE
  - 8 per host/128 per datastore if using 10GbE
  - ...and that's before you tweak settings for more, and shoot yourself in the foot :-)
- **Need to meet the vMotion network requirements**
  - 622Mbps or more, 5ms RTT (upped to 10ms RTT if using Metro vMotion - vSphere 5 Enterprise Plus)
  - Layer 2 equivalence for vmkernel (support requirement)
  - Layer 2 equivalence for VM network traffic (required)

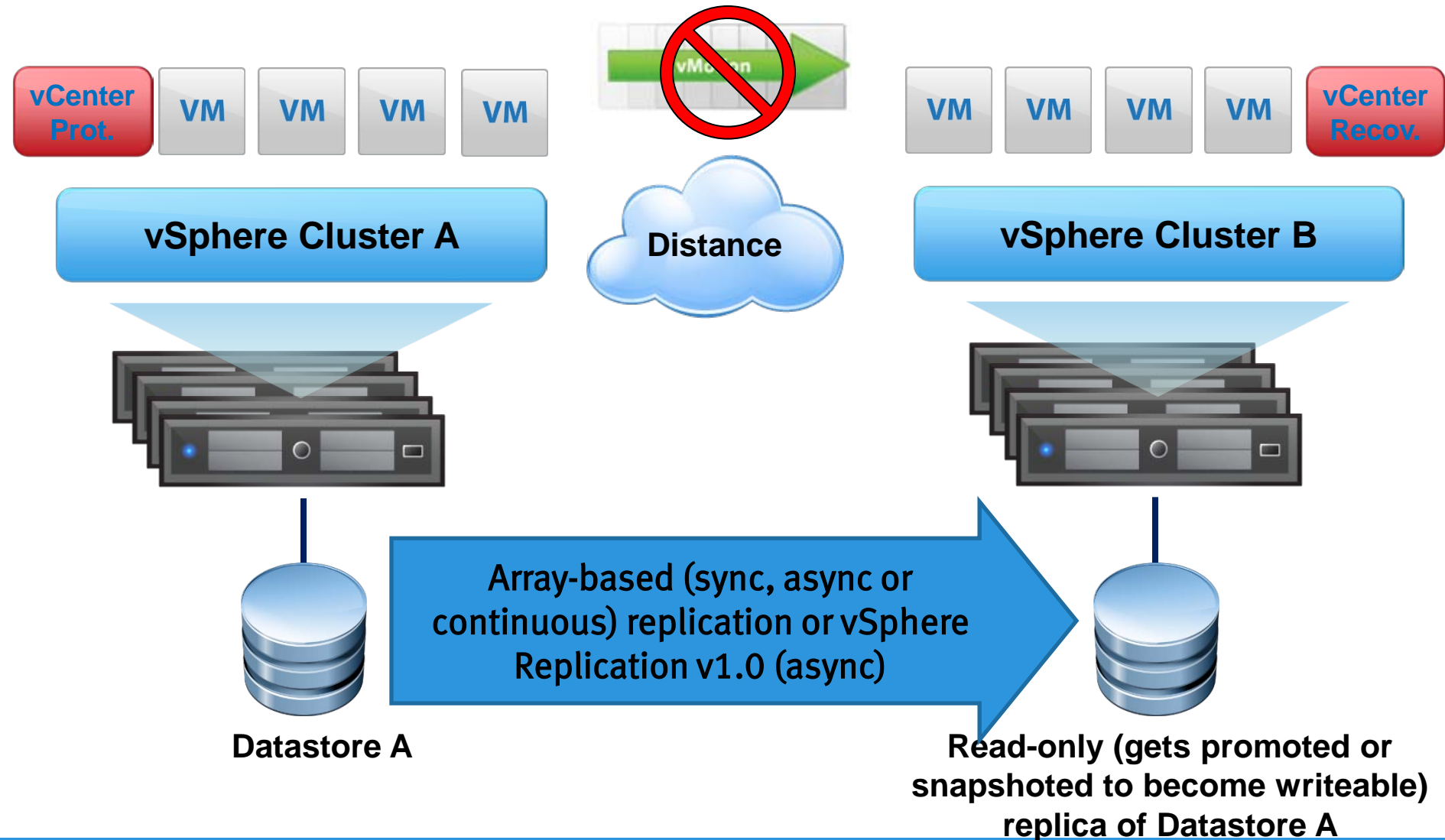
# Type 2: “Multiple vSphere Clusters”



# One little note re: “Inter-Cluster” vMotion

- **Inter-Cluster vMotions are serialized**
  - Involves additional calls into vCenter, so hard limit
  - Lose VM cluster properties (HA restart priority, DRS settings, etc.)
- **Need to meet the vMotion network requirements**
  - 622Mbps or more, 5ms RTT (upped to 10ms RTT if using Metro vMotion w vSphere 5 Enterprise Plus)
  - Layer 2 equivalence for vmkernel (support requirement)
  - Layer 2 equivalence for VM network traffic (required)

# Type 3: “Classic Site Recovery Manager”



# Part I - Summary

- People have a hard time with this... Disaster Avoidance != Disaster Recovery
  - Same logic applies at a server level applies at the site level
  - Same value (non-disruptive for avoidance, automation/simplicity for recovery) that applies at a server level, applies at the site level
- Stretched clusters have many complex considerations
- SRM and non-disruptive workload mobility are mutually exclusive right now
  - vMotion = single vCenter domain vs. SRM = two or more vCenter domains
  - *Note – people use SRM for workload mobility all the time (and is improved in vSphere 5/SRM 5) – but this is always disruptive*
  - **SRM remains the simplest, cleanest solution across the majority of use cases**
  - **But, there are use cases where stretched vSphere clusters have a place**

# PART II...

## vSphere Stretched Clusters Considerations

# Stretched Cluster Design Considerations

- **Understand the difference compared to DR**
  - HA does not follow a recovery plan workflow
  - HA is not site aware for applications, where are all the moving parts of my app? Same site or dispersed? How will I know what needs to be recovered?
- **Single stretch site = single vCenter**
  - During disaster, what about vCenter setting consistency across sites? (DRS Affinity, cluster settings, network)
- **Will network support? Layer2 stretch? IP mobility?**
- **Cluster split brain (at both the storage and VMware layers) = big concern, how to handle?**

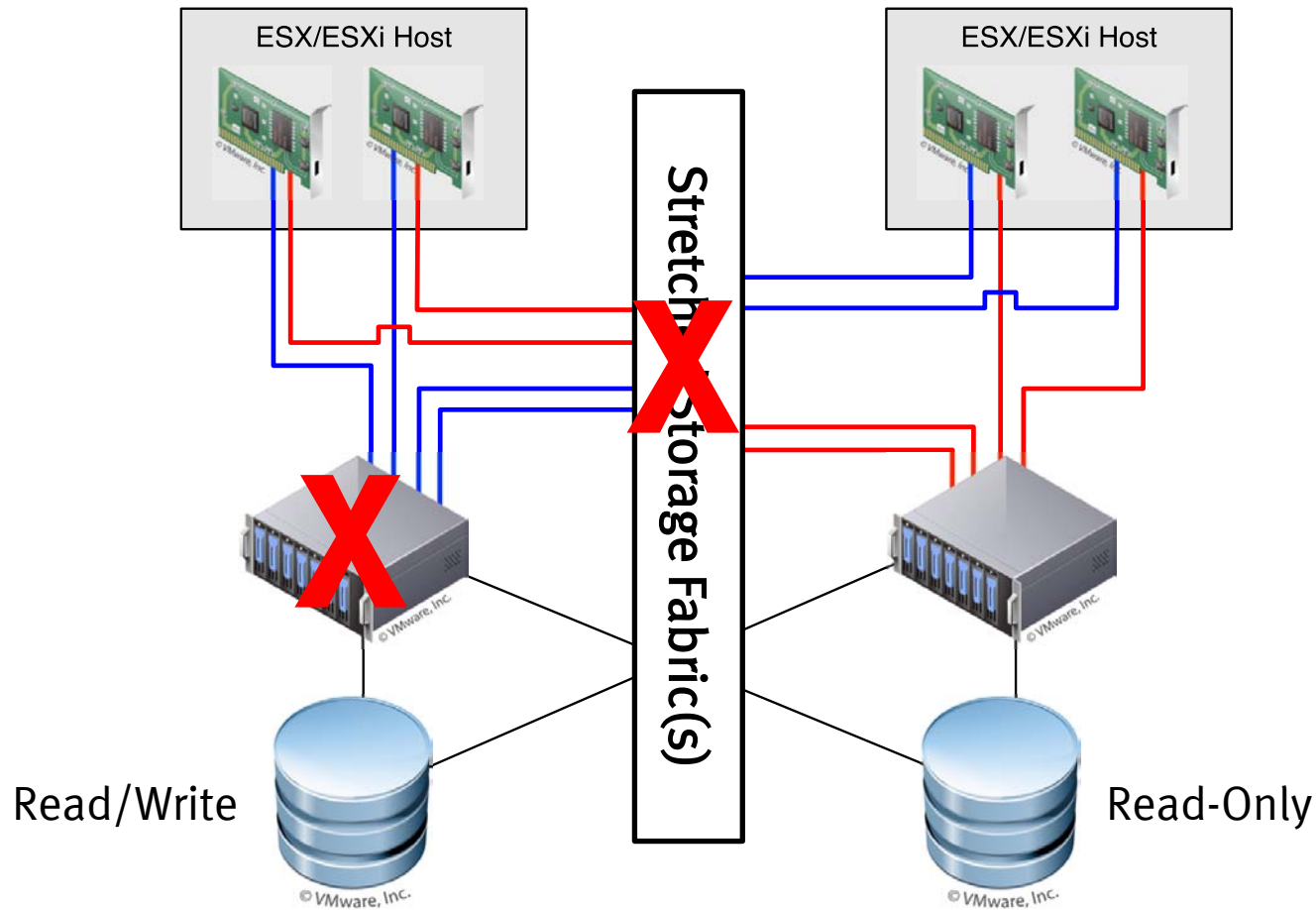
*Not necessarily cheaper solution, read between the lines (there is usually hidden storage, networking and WAN costs)*

# “Uniform Access” Stretched Storage Model

- Characterized by all the ESX hosts in a cluster accessing a storage device at the same IQN, NFS server, or WWN
- Literally just stretching the SAN fabric or LAN when using NFS between locations
- Requires synchronous replication (to recovery on partition)
- Limited in distance to ~100km in most cases
- Read/write in one location, read-only in second location
- Implementations with only a single storage controller at each location create other considerations
- Must address storage layer “split brain” scenarios



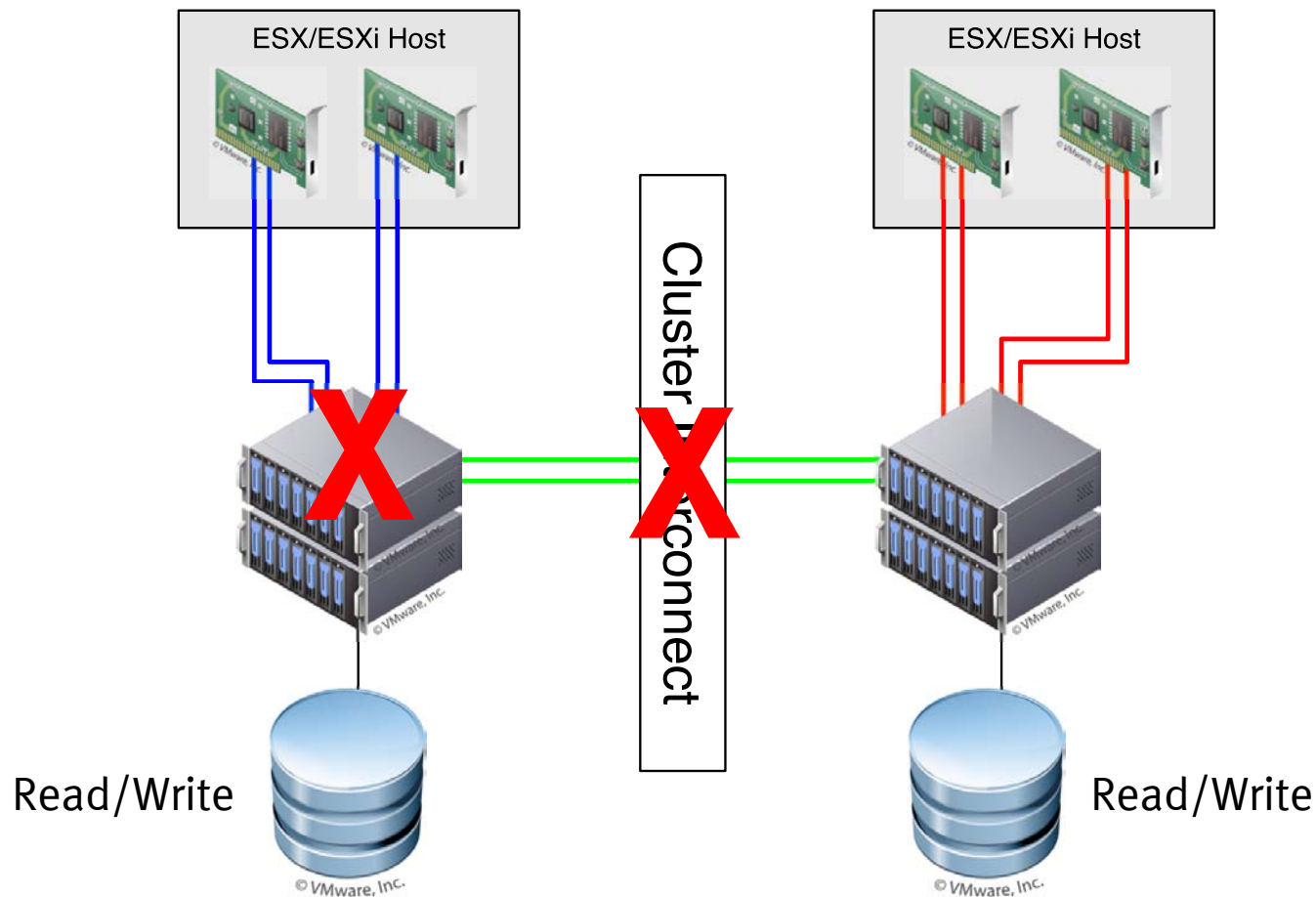
# “Uniform Access” Stretched Storage Model



# “Non Uniform Access” Stretched Storage Model

- Characterized by the ESX hosts on each side accessing a different storage device via IQN, NFS server, or WWN
- Leverages new storage technologies to distribute storage across multiple sites
- TODAY - requires synchronous mirroring
- Limited in distance to ~100km in most cases
- Read/write storage in both locations, employs data locality algorithms
- Typically uses multiple controllers in a scale-out fashion
- Must address storage layer “split brain” scenarios

# “Non Uniform Access” Stretched Storage Model



# EMC VPLEX Overview

- EMC VPLEX falls into the “Non Uniform Access” Stretched Storage category
- Keeps data synchronized between two locations but provides read/write storage simultaneously at both locations
- Uses scale-out architecture with multiple engines in a cluster and two clusters in a Metro-Plex
- Supports both EMC and non-EMC arrays behind the VPLEX

# VPLEX – What A Metro-Plex looks like

The screenshot displays the EMC VPLEX Management Console V4.2 interface. The browser address bar shows the URL `https://10.12.177.189/smsflex/VPlexConsole.html`. The console has a navigation bar with tabs: System Status (selected), Provisioning Overview, Provision Storage, Mobility Central, and Help. The main content area is titled "System Status: EMC VPLEX™ Metro" and includes a sub-header "Get a quick overall status and a summary of components in each cluster of your Metro-Plex. [Learn more...](#)".

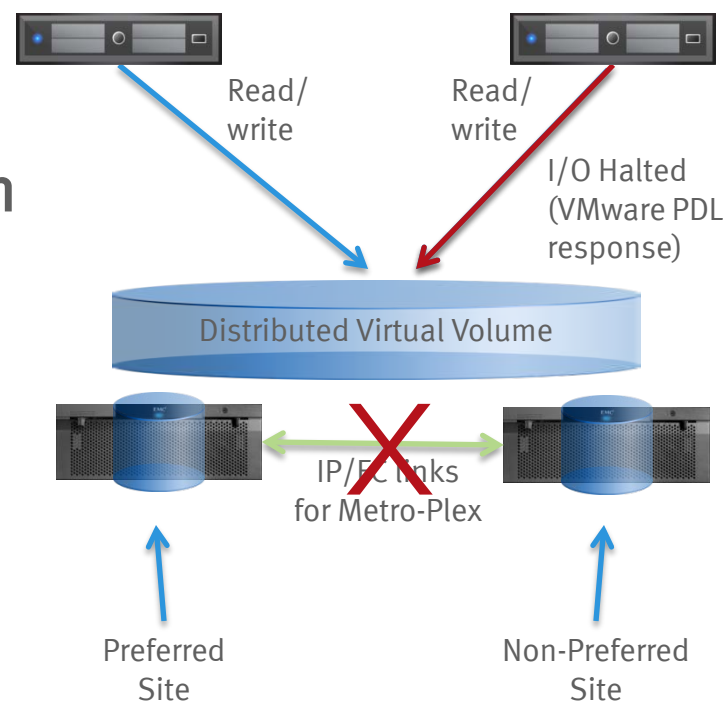
Two clusters are shown:

- cluster-1**:
  - Storage Views: 3 (ok)
  - Directors: 2 (ok)
  - Arrays: 2 (degraded)
  - Operational Status:
  - Health:
- cluster-2**:
  - Storage Views: 1 (stopped), 3 (ok)
  - Directors: 2 (ok)
  - Arrays: 1 (ok)
  - Operational Status:
  - Health:

The bottom of the console shows the user "service" and a status bar with "Done", "Internet", and "100%" indicators.

# Preferred Site in VPLEX Metro

- VPLEX Metro provides read/write storage in two locations at the same time (AccessAnywhere)
- In a failure scenario, VPLEX uses “detach rules” to prevent split brain
  - A preferred site is defined on a per-distributed virtual volume (not site wide) basis
  - Preferred site remains read/write; I/O halted at non-preferred site
- Invoked only by entire cluster failure, entire site failure, or cluster partition



# Configuring Preferred Site...

The screenshot displays the EMC VPLEX Management Console V4.2 interface. The main window shows the 'Provision Storage' section with a tree view on the left containing 'Distributed Storage', 'Distributed Devices', 'Clusters', 'Hosts', 'Storage Views', 'Initiators', 'Ports', 'Virtualized Storage', 'Virtual Volume', 'Devices', 'Extents', 'Physical Storage', 'Storage Volum', and 'Storage Array'. The 'Distributed Devices' table lists several devices, including 'DR1\_device1', 'DR1\_device2', 'DR1\_io\_dr', and various 'device\_2qdr' entries. A 'Distributed Device Properties' dialog box is open, showing the 'Distributed Device Name' as 'DR1\_device1', 'Transfer Capacity' as '46.2G', and 'Virtual Volume' as 'DR1\_device1\_vol'. The 'Rule Set' dropdown menu is highlighted with a red box and set to 'cluster-2-detaches'. Other properties shown include 'Health: ok', 'Health Indications: None', 'Operational Status: ok', and 'Service Status: running'. The 'Components of Selected Device' panel on the right shows a list of components for 'DR1\_device1', including 'cluster-1', 'dr\_device\_c1', and 'cluster-2', each with a list of 'extent' entries.

# Something to understand re: yanking & “suspending” storage...

- What happens when you “yank” storage?
  - VMs who’s storage “disappears” or goes “read-only” behave *indeterminately*
  - Responding to a ping doesn’t mean a system is available (if it doesn’t respond to any services, for example)
  - There’s no chance of “split brain” data
  - But – VMs can stay alive for surprisingly long
  - Conversely, sometimes, VMs blue-screen quickly
- Yanked: <http://www.youtube.com/watch?v=6Op0i0cekLg>
- Suspended: <http://www.youtube.com/watch?v=WJQfy7-udOY>



# Stretched Cluster Considerations #1

**Consideration:** Without read/write storage at both sites, roughly half the VMs incur a storage performance penalty on every IO

- **With stretched “Uniform Access” configurations:**
  - VMs running in one site are accessing storage in another site
  - Creates additional latency for every I/O operation
- **With “Non-Uniform” configurations:**
  - Read/write storage provided, so this doesn’t apply

# Stretched Cluster Considerations #2

Consideration: Prior to and including vSphere 4.1, you CAN NOT control HA/DRS behavior for “sidedness” – and this is material on VM latency

- **With stretched Storage Network configurations:**
  - Additional latency introduced when VM storage resides in other location
  - Storage vMotion required to remove this latency
- **With distributed virtual storage configurations:**
  - Need to keep cluster behaviors in mind
  - Data is access locally due to data locality algorithms

# Stretched Cluster Considerations #3

Consideration: With vSphere 5, you CAN use DRS host affinity rules to control HA/DRS behavior

- With all storage configurations:
  - Not automatic (so... Do you have operational discipline?)
- With “Uniform Access” configurations:
  - Beware of single-controller implementations
  - Storage latency still present in the event of a controller failure
- With “Non-Uniform Access” configurations:
  - Plan for cluster failure/cluster partition behaviors
- *Worth re-iterating – host affinity in stretched cluster use cases not supported in vSphere 4.1. More on vSphere 5 in a bit.*

# Stretched Cluster Considerations #4

Consideration: There is no supported way to control VMware HA primary /secondary node selection with vSphere 4.x

- **With all storage configurations:**
  - Limits cluster size to 8 hosts (4 in each site)
  - No **supported** mechanism for controlling/specifying primary/secondary node selection
  - Methods for increasing the number of primary nodes also not supported by VMware
- *Note: highly recommended reading (just ignore non-supported notes) : <http://www.yellow-bricks.com/vmware-high-availability-deepdiv/>*
- *More on vSphere 5 in a bit.. .*

# Stretched Cluster Considerations #5

**Consideration:** *Stretched HA/DRS clusters (and inter-cluster vMotion also) require Layer 2 “equivalence” at the network layer*

- With all storage configurations:
  - Complicates the network infrastructure
  - Involves technologies like OTV, VPLS/Layer 2 VPNs
- VXLAN may help with this
  - But, is a tech preview only at this point
- *Note how the SRM automated IP change is much simpler in many cases*

# Stretched Cluster Considerations #6

Consideration: The network lacks site awareness, so stretched clusters introduce new networking challenges

- With all storage configurations:
  - The movement of VMs from one site to another doesn't update the network
  - VM movement causes “horseshoe routing” (LISP, a future networking standard, helps address this)
  - You'll need to use multiple isolation addresses in your VMware HA configuration
- *Note how the SRM automated IP change is much simpler in many cases*
- *Highlights – pursue stretched clusters only if you have networking expertise in house*

# Summary – and recommendations

Don't let storage vendors (me included) do the Jedi mind trick on you.

*In the paraphrased words of Yoda... “think not of the sexy demo, think of operations during the disaster – there is no try”*

# PART III...

What's new...



# So – what's new?

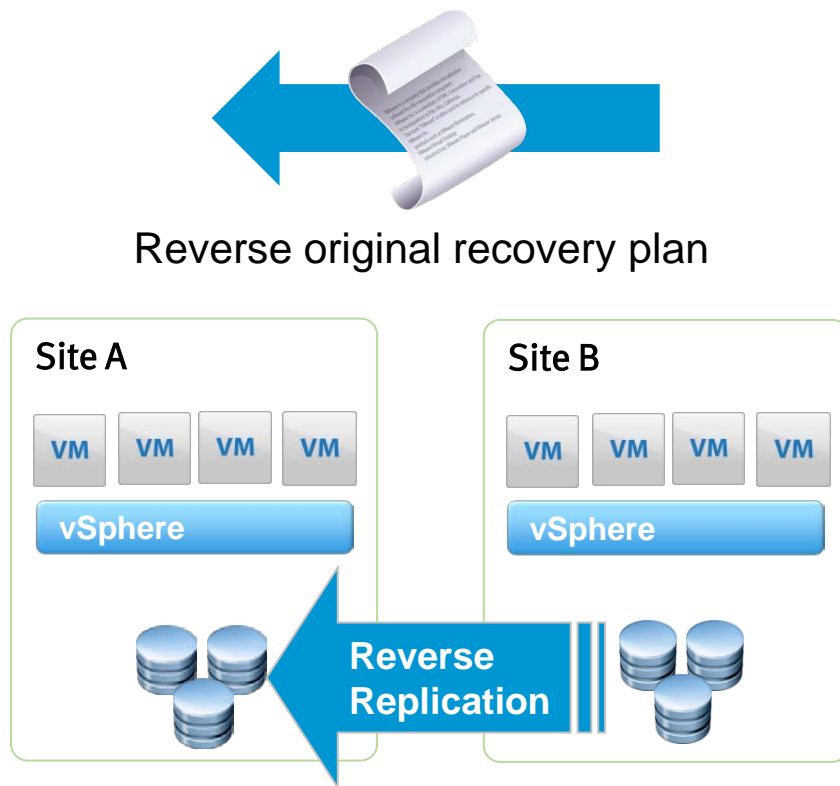
- **NOW** – Site Recovery Manager 5.0
- **NOW** – vSphere 5 VM HA rewrite & heartbeat datastores, help on partition scenarios
- **NOW** – vSphere 5 Metro vMotion
- **NOW** – Improved VPLEX partition behavior – will mark the target as “dead”, works better with vSphere
- **NOW** – VPLEX cluster interconnect and 3<sup>rd</sup> party witness

# SRM 5.0 New Features

- New Workflows – inc Failback!!!
- Planned migration – with replication update
- vSphere Replication framework
- Redesigned UI – true single pane of glass configuration
- Faster IP customization
- SRM specific Shadow VM icons at recovery site
- In guest scripts callout via recovery plans
- VM dependency ordering during configurable
- ...and a LOT more...

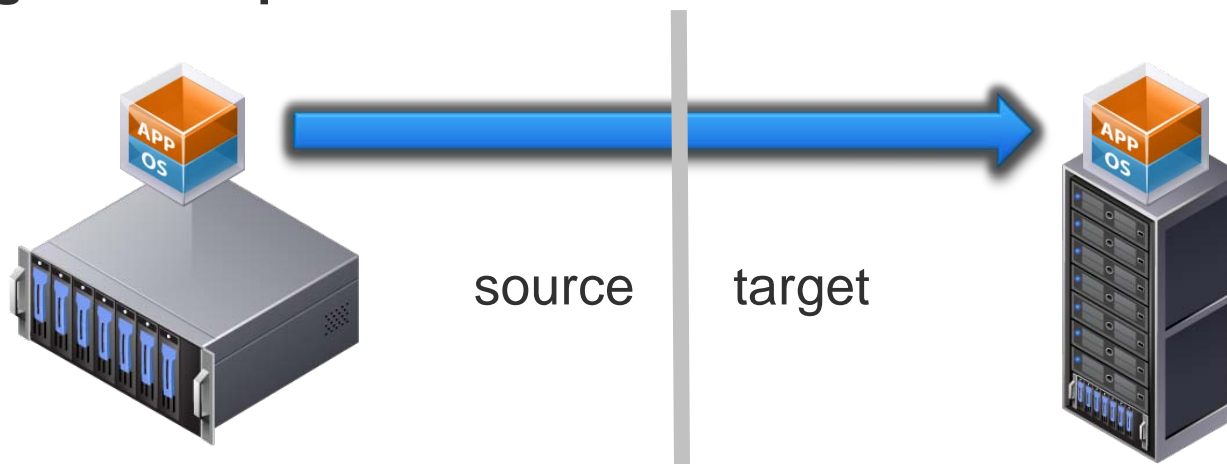
# SRM 5.0 – Automated Failback

- **Reprotect VMs from Site B to Site A**
  - Reverse Replication
  - Apply reverse resource map
- **Automate failover Site B to Site A**
  - Reverse original recovery plan
- **Simplify failback process**
  - Automate replication management
  - Eliminate need to set up new recovery plan and cleanup
- **Restrictions**
  - Does not apply if Site A physically lost
  - Not available at GA with vSphere Replication



# SRM 5.0 – vSphere Replication

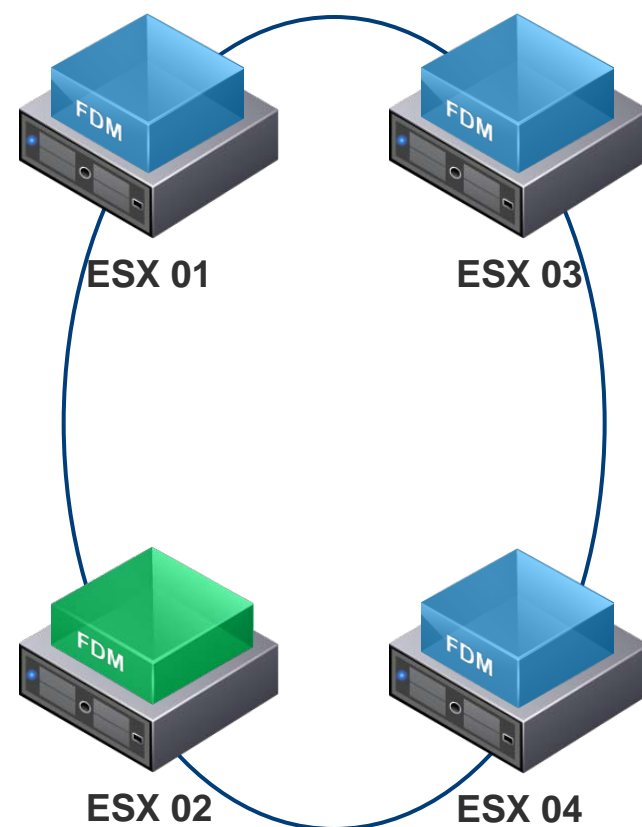
- Adding native replication to SRM



- VMs can be replicated **regardless** of the underlying storage
- Enables replication between **heterogeneous** datastores
- Replication is managed as a **property** of a virtual machine
- Efficient replication **minimizes** impact on VM workloads
- Considerations: Scale, Failback, Consistency Groups

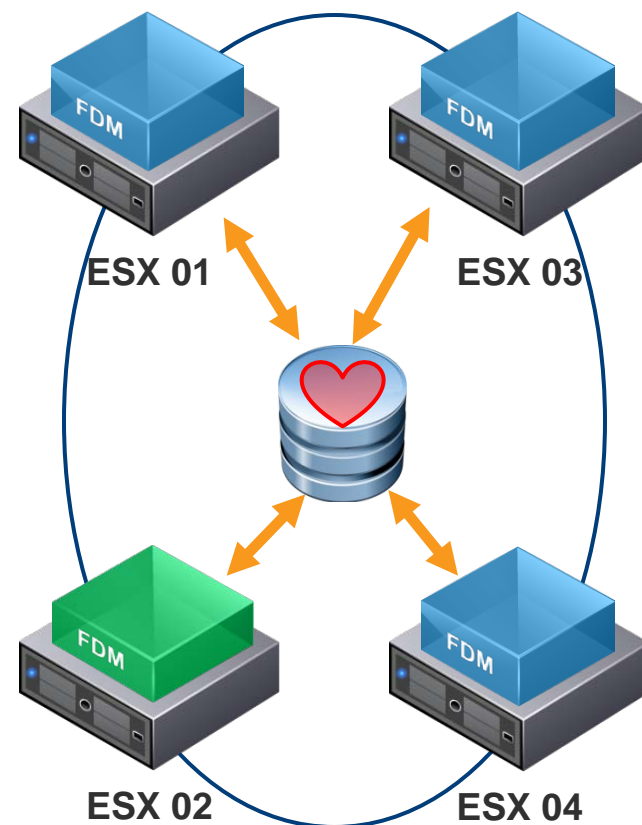
# vSphere 5.0 - HA

- Complete re-write of vSphere HA
- Elimination of Primary/Secondary concept
- Foundation for increased scale and functionality
  - Eliminates common issues (DNS resolution)
- Multiple Communication Paths
  - Can leverage storage as well as the mgmt network for communications
  - Enhances the ability to detect certain types of failures and provides redundancy
- IPv6 Support
- Enhanced User Interface
- Enhanced Deployment



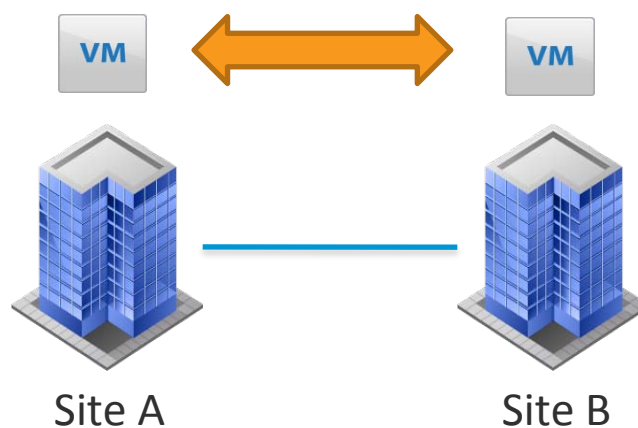
# vSphere 5.0 HA – Heartbeat Datastores

- Monitor availability of Slave hosts and VMs running on them
- Determine host network isolated VS network partitioned
- Coordinate with other Masters – VM can only be owned by one master
- By default, vCenter will automatically pick 2 datastores
- *Very important for hardening stretched storage models on partition*



# Metro vMotion – Stretched Clusters

- Feature when vSphere 5 Enterprise Edition is licensed
- Enables vMotion across longer distances through changes in the vmkernel stack
- Facilitates workload balancing between sites
- vMotion is less latency sensitive (up to 10ms RTT)



# ...All Culminating in the vSphere Metro Stretch Cluster (vMSC) HCL!

- For the first time, VMware QA's stretched cluster use cases
- For the first time – with vSphere 5 this means that VMware formally supports some stretched cluster configurations
- Previously (vSphere 4.1) configurations work, have KB articles, and the storage vendor is responsible for support
- New comprehensive automated test harness – built and used collaboratively – *this will accelerate adoption of vSphere Metro Stretched Cluster use cases*



# See more here...

The image shows two overlapping web pages from VMware. The background page is the 'VMware Compatibility Guide' with a search bar and filters for Product Release Version, Array Type, Partner Name, and Keyword. The foreground page is the 'VMware Knowledge Base' for the article 'Implementing vSphere Metro Storage Cluster (vMSC) using EMC VPLEX'. The article includes sections for Purpose, Resolution, What is VPLEX?, and What is vMSC?. A sidebar on the right contains action links like 'Bookmark Document', 'Copy URL', and 'Email Document', along with a rating of 9 stars and the article ID 2007545.

Home > Resources > Compatibility Guides

## VMware Compatibility Guide

Search Compatibility Guide: ? (e.g. compatibility or esx or 3.0)

Looking for management tools optimized for ESXi? [See the partner page](#)

What are you looking for: Storage/SAN

**Product Release Version:**

- All
- ESXi 5.0
- ESX 4.1 U1
- ESX 4.1
- ESX 4.0 U3
- ESX 4.0 U2

**Array Type:**

- All
- FCoE
- FC
- iSCSI

**Storage Virtual Appliance**

Only: ☐ Yes ☒ No

**Partner Name:**

- All
- 3PAR
- Aberdeen LLC
- AC&NC

**Firmware Version:**

- All
- FLARE 02.19.500.5.1
- M110R21

**Keyword:**

**Posted Date Range:**

All

[Update and View Results](#) [Reset](#)

vmware®

Community | Forums | Technical Resources | Virtual Appliances | Store | My Account

Cloud Computing | Virtualization | Solutions | Products | Services | Support & Downloads | Partners | Company

## Knowledge Base

The VMware Knowledge Base provides support solutions, error messages and troubleshooting guides

KB Home Knowledge Base Help

Search the VMware Knowledge Base (KB)

Search

Products --> Category -->

View by Article ID  View

### Implementing vSphere Metro Storage Cluster (vMSC) using EMC VPLEX

★★★★★  
9 Ratings

**Actions**

- Bookmark Document
- Copy URL
- Email Document
- Print Document
- Subscribe to Document
- SHARE

1 tweet

retweet

KB Article: 2007545

Updated: Oct 11, 2011

Categories: Informational

Products: VMware ESXi

**Purpose**

This article provides information about deploying a Metro Storage Cluster across two datacenters using EMC VPLEX Metro 5.0. With vSphere 5.0, a Storage Virtualization Device can be supported in a Metro Storage cluster configuration.

**Resolution**

**What is VPLEX?**

EMC VPLEX is a federation solution that provides simultaneous access to storage devices at two geographically separate sites. One or more VPLEX Distributed Virtual Volumes can be provisioned for sharing between the two sites' ESXi hosts. These volumes can be used as Raw Device Mapping (RDM) disks or as a shared VMFS datastore. The RDM can be used for exclusive access by the virtual machine and the VMFS datastore can be used for provisioning virtual machines and carving out additional vDisks.

The VPLEX cluster at each site itself is designed to be highly available. A VPLEX cluster can scale from two directors to eight directors. Each director is protected by redundant power supplies, fans, and interconnects, making the VPLEX highly resilient.

**What is vMSC?**

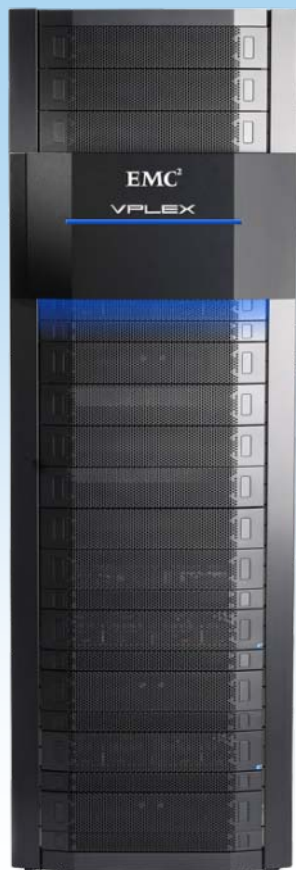
vSphere Metro Storage Cluster (vMSC) is new configuration. A storage device configured in the MSC configuration is



Imp vMSC EMC  
VPLEX.pdf

EMC<sup>2</sup>

# What's new with VPLEX 5.0



NEW

EMC²

ORACLE



DELL



NetApp

IBM

FUJITSU

HITACHI

NEW

## GeoSynchrony 5.0 for VPLEX

- Expanded 3<sup>rd</sup> party storage support
- VP-Copy for EMC arrays
- Expanded array qualifications (ALUA)
- VPLEX Witness
- Host cross-cluster connected
- VPLEX Element Manager API
- VPLEX Geo

EMC²

# VS2: New VPLEX Hardware



- Faster Intel multi-core processors
- Faster engine interconnect interfaces
- Space-efficient engine form factor



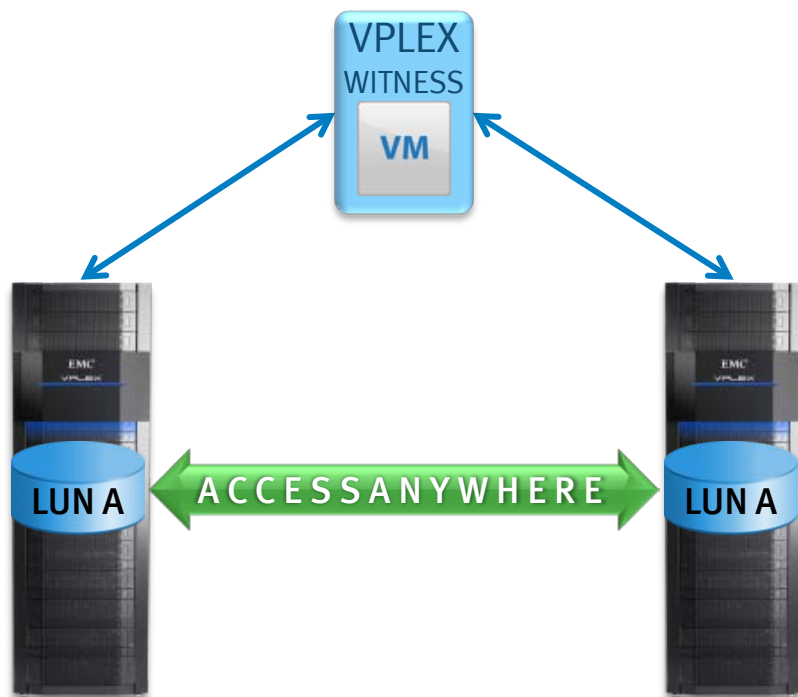
- Third-party rack support

Migrated an entire datacenter – saving \$500,000 in revenue VPLEX paid for itself twice-over in a single event. As a hospital they did not have to interrupt healthcare.

“I'm sure glad we made the DR investment. It took a lot of pressure off us. We ran the DR virtual farm over 50 hours. This is solid stuff. VPLEX is well worth the investment by the way.”

CIO, Northern Hospital of Surry County

# VPLEX Witness

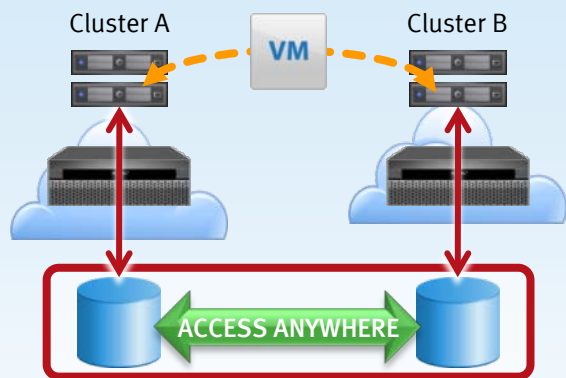


- Use with VPLEX Metro and VPLEX Geo
- Coordinates seamless failover
- Runs as a virtual machine within an ESX host
- Connects to VPLEX through IP

Integrates with hosts, clusters, applications to automate failover and recovery

# VPLEX Family Use Cases

## MOBILITY



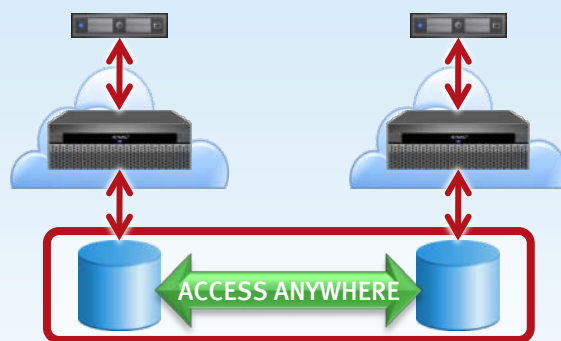
Move and relocate VMs, applications, and data over distance

**Disaster avoidance**

**Data center migration**

**Workload rebalancing**

## AVAILABILITY

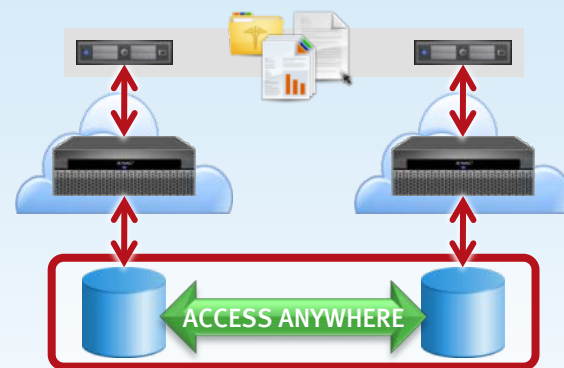


Maintain availability and non-stop access by mirroring across locations

**High availability**

**Eliminate storage operations from failover**

## COLLABORATION



Enable concurrent read/write access to data across locations

**Instant and simultaneous data access over distance**

**Streamline workflow**

# VPLEX Family Product Matrix

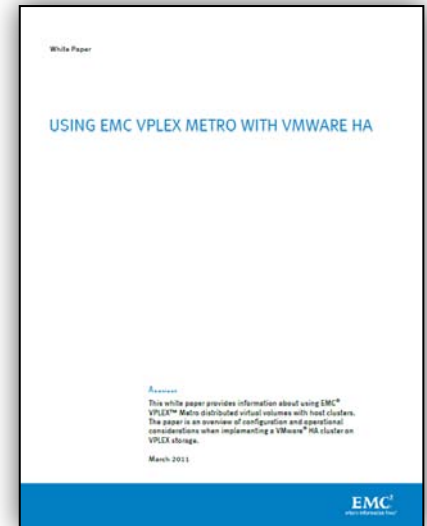
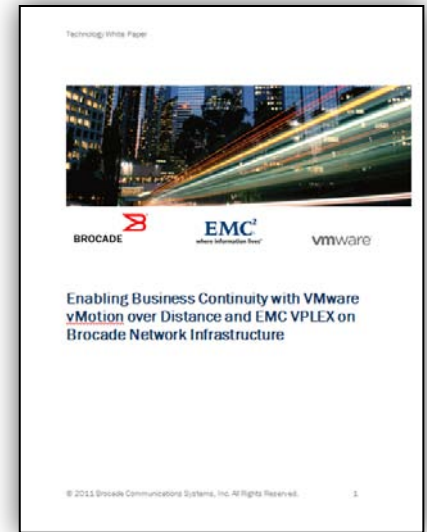
## VPLEX

	Local	Metro	Geo
<b>Mobility</b> Within a data center	✓	✓	✓
Synchronous: approximately 100 km		✓	
Asynchronous: approximately 1,000 km			✓
<b>Availability</b> High availability	✓	✓	✓
VPLEX Witness support		✓	✓
Cross-cluster connected configuration		✓	
<b>Collaboration</b> Between two sites		✓	✓

EMC<sup>2</sup>

# For More Information...

- vSphere Metro Stretched Cluster (VMware supported) KB:
  - <http://kb.vmware.com/kb/2007545>
- vSphere 4.1 (remember, all stretch clusters are vendor supported, not VMware supported)
  - Using VPLEX Metro with VMware HA:  
<http://kb.vmware.com/kb/1026692>
  - vMotion over Distance Support with VPLEX Metro:  
<http://kb.vmware.com/kb/1021215>
- VPLEX Metro HA techbook  
[http://powerlink.emc.com/km/live1/en\\_US/Offering\\_Technical/Technical\\_Documentation/h7113-vplex-architecture-deployment-techbook.pdf](http://powerlink.emc.com/km/live1/en_US/Offering_Technical/Technical_Documentation/h7113-vplex-architecture-deployment-techbook.pdf)
- VPLEX Metro with VMware HA  
[http://powerlink.emc.com/km/live1/en\\_US/Offering\\_Technical/White\\_Paper/h8218-vplex-metro-vmware-ha-wp.pdf](http://powerlink.emc.com/km/live1/en_US/Offering_Technical/White_Paper/h8218-vplex-metro-vmware-ha-wp.pdf)



# PART IV...

What we're working on...

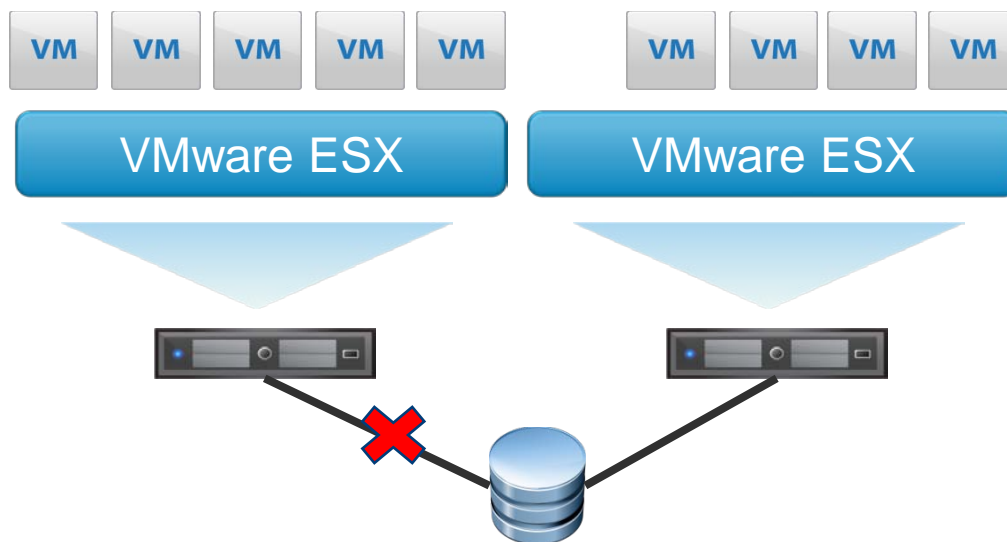


# The coolness is accelerating...

- Ongoing SRM and VM HA enhancements
- Expanding “side awareness” in vSphere
- Improving stretched cluster + SRM coexistence
- Too many things to cover today... quick look at just a few of them...

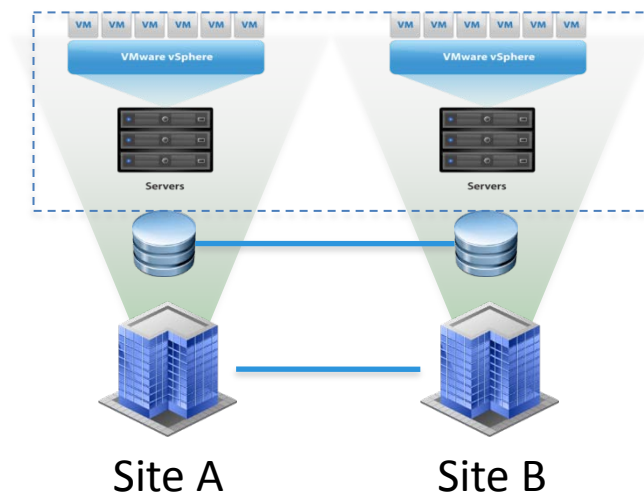
# VM Component Protection

- Detect and recover from catastrophic infrastructure failures affecting a VM
  - Loss of storage path
  - Loss of Network link connectivity
- VMware HA restarts VM on available healthy host



# Automated Stretched Cluster Config

- Leverage the work in VASA and VM Granular Storage (VSP3205)
- Automated site protection for all VM's
- Benefits of single cluster model
- Automated setup of HA and DRS affinity rules

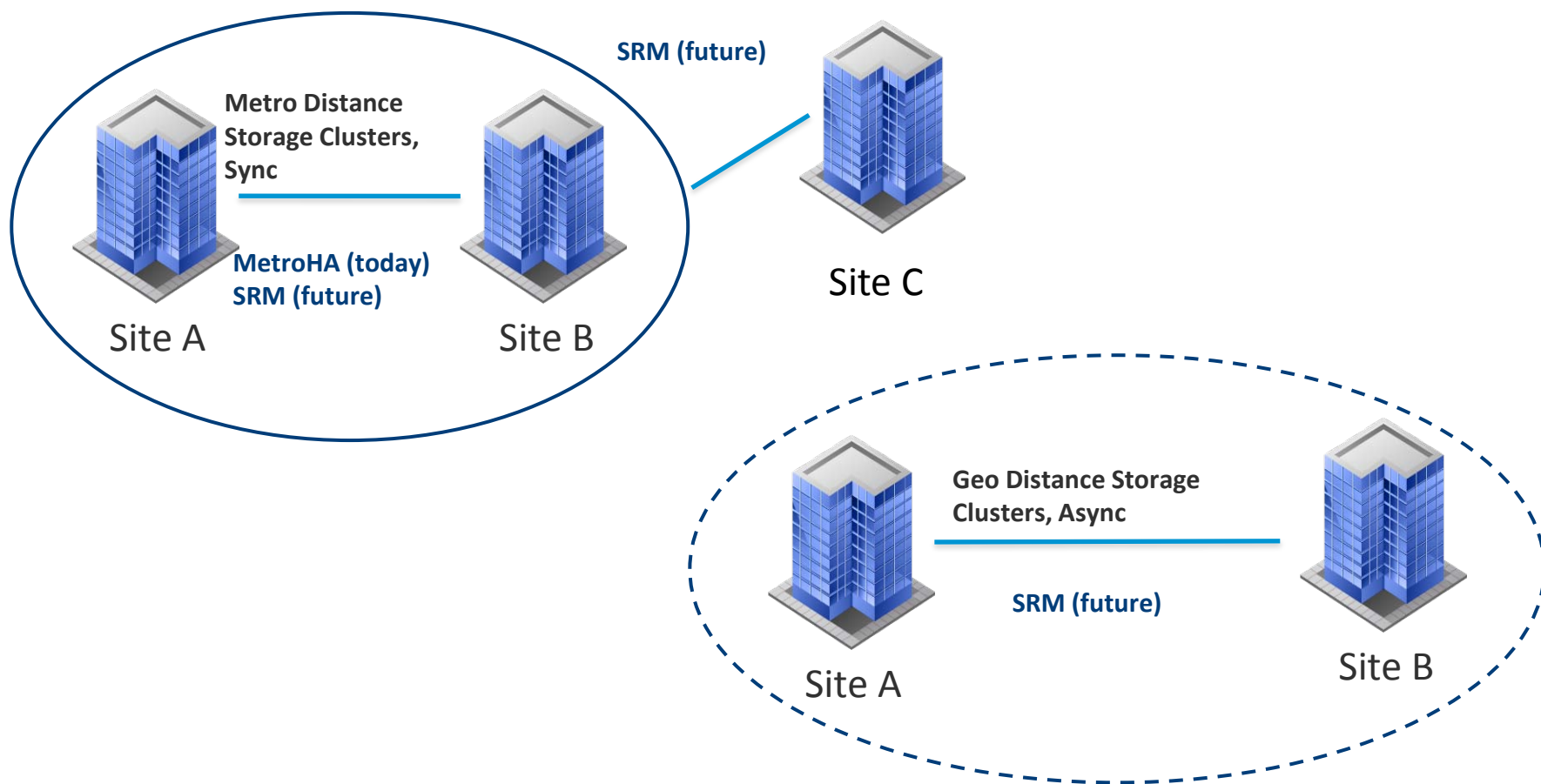


HA/DRS Cluster

Distributed Storage Volumes

Layer 2 Network

# Increased Topology Support



# Q & A – Part 1 – Questions from us to you

- “I think a stretched cluster is what we need... How do I know?”
- “I think a DR solution is what we need... How do I know?”
- “Stretched clustering sounds like awesomesauce, why not?”
- “Our storage vendor/team tells us their disaster avoidance solution will do everything we want, HA, DA, DR, we are not experts here, should we be wary?”
- “Our corporate SLA’s for recovery are simple BUT we have LOTS of expertise and think we can handle the bleeding edge stuff should we just go for it???”

## Q & A – Part 2 – Questions from us to you

- “Can we have our cake and eat it yet? We want BOTH solutions together?”
- “Is there anything the storage vendors are NOT telling us that might make running this day to day costly from an opex point of view?”
- “Why does one solution use a single vCenter yet the other uses two?? the DR solution seems less flexible and more complex to manage, is that fair?”
- “My datacenter server rooms are 50 ft apart but i definitely want a DR solution what's wrong with that idea?”

# Q & A – Part 3 – We would love to hear...

## Looking to async distances...

- Is “cold migration” over distance good enough for you, or is it live or nothing?
- Would you pay for it (easiest litmus test of “nice to have”)
- Would you be willing to be very heterogeneous to use it?
- What are your thoughts on networking solutions (are you looking at OTV type stuff?)

THANK YOU