

VMware云资源池业务连续性设计

刘承罡

13581900001

Confidential

vmware®

什么是云资源池的业务连续性？

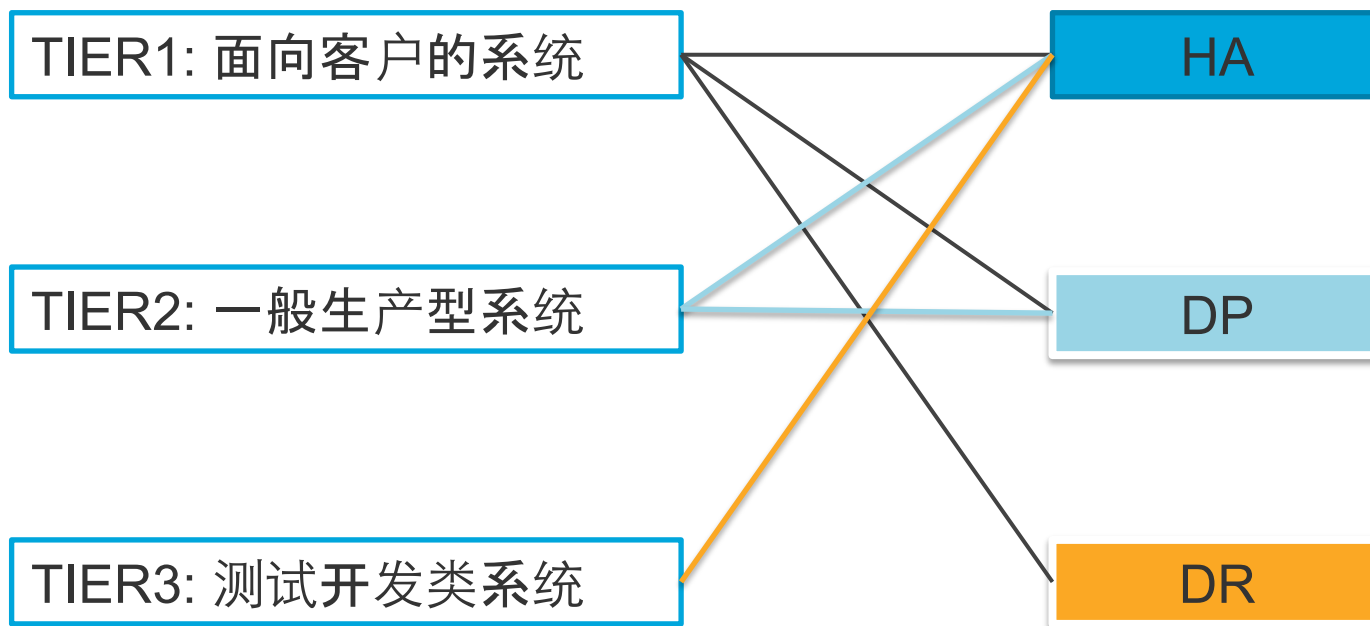
IT Business Continuity



理解DR和DA的区别

- **Disaster Recovery (DR)** – 在灾难发生后的一整套恢复程序,一般都会恢复一份过去数据的完整的、一致的数据和业务拷贝。发生在业务中断之后,有服务、OS重启操作,最终用户有感知的,有RTO/RPO的概念: HA/DP/SRM
- **Disaster Avoidance (DA)**- 在灾难发生之前进行预测,并提前预防防止业务中断,根据技术选择的不同,用户一般无感知,无需用户业务、OS的重启操作,典型的DA应用如: vMotion

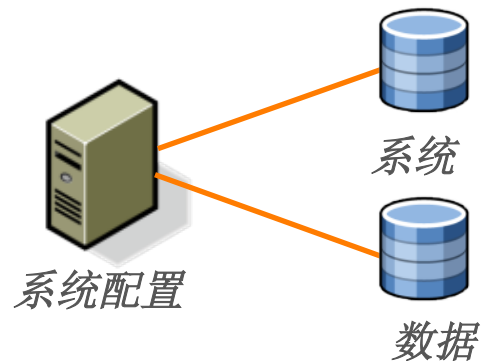
按照业务的QOS等级实现设计业务连续性方案



传统物理机与虚拟机环境对系统和数据的保护不同

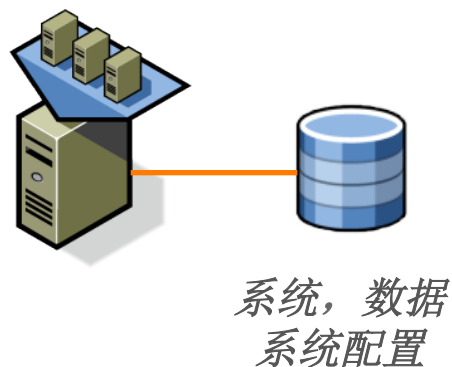
物理机环境中系统和数据的保护

- 系统磁盘和数据采用不同的方式来保护
- 为了确保恢复的成功率，一般要求采用相等的硬件设施
- 需要非常复杂的流程才能实现保护成功



虚拟架构下系统和数据的保护

- 系统磁盘和数据采用同样的方式来保护
- 系统也是数据
- 保护系统的同时就保护了数据，反之亦然



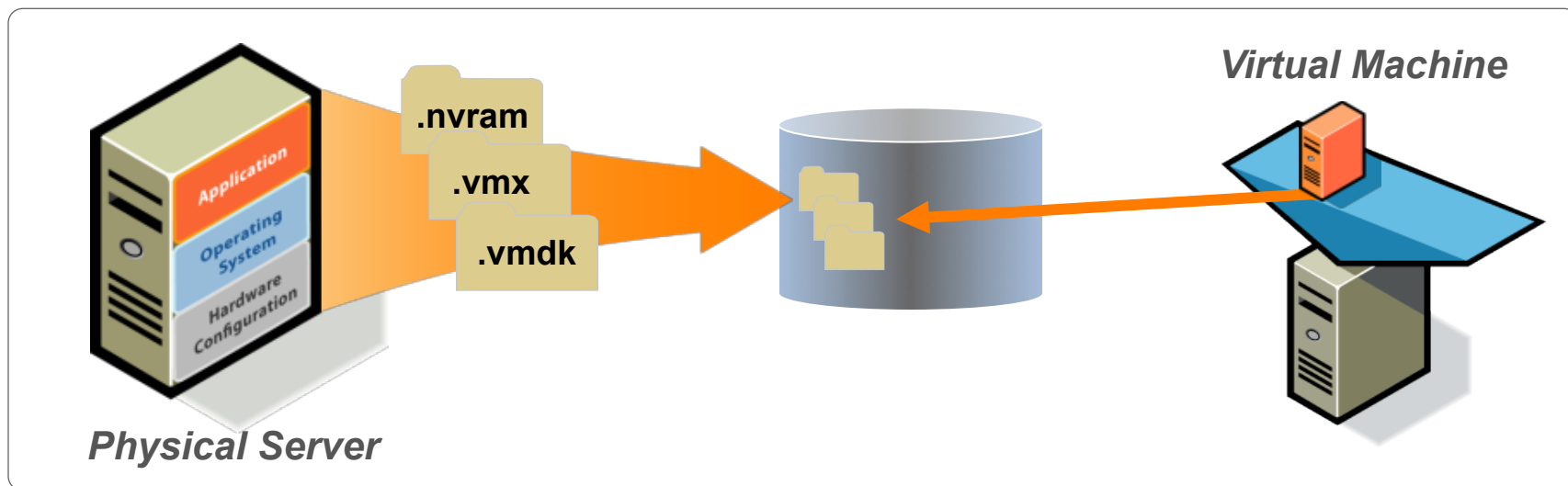
虚拟架构是化繁为简的根本原因

虚拟机既存储系统也存储数据

- 整个系统封装在文件中：
硬件配置，操作系统，应用及数据

影响

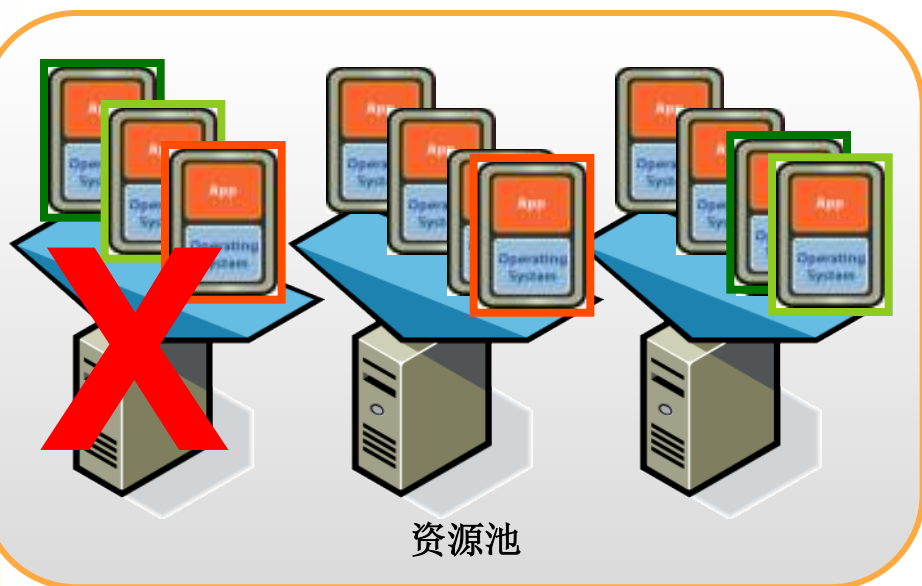
- 系统即数据
 - 使用保护数据的同样的方法和流程来保护系统
- 虚拟机是存储系统最简单灵活的方法



云资源池下的高可用

通过VMware HA确保系统高可用

通过VMware HA可以方便地提高任何应用的高可用性



VMware HA是什么？

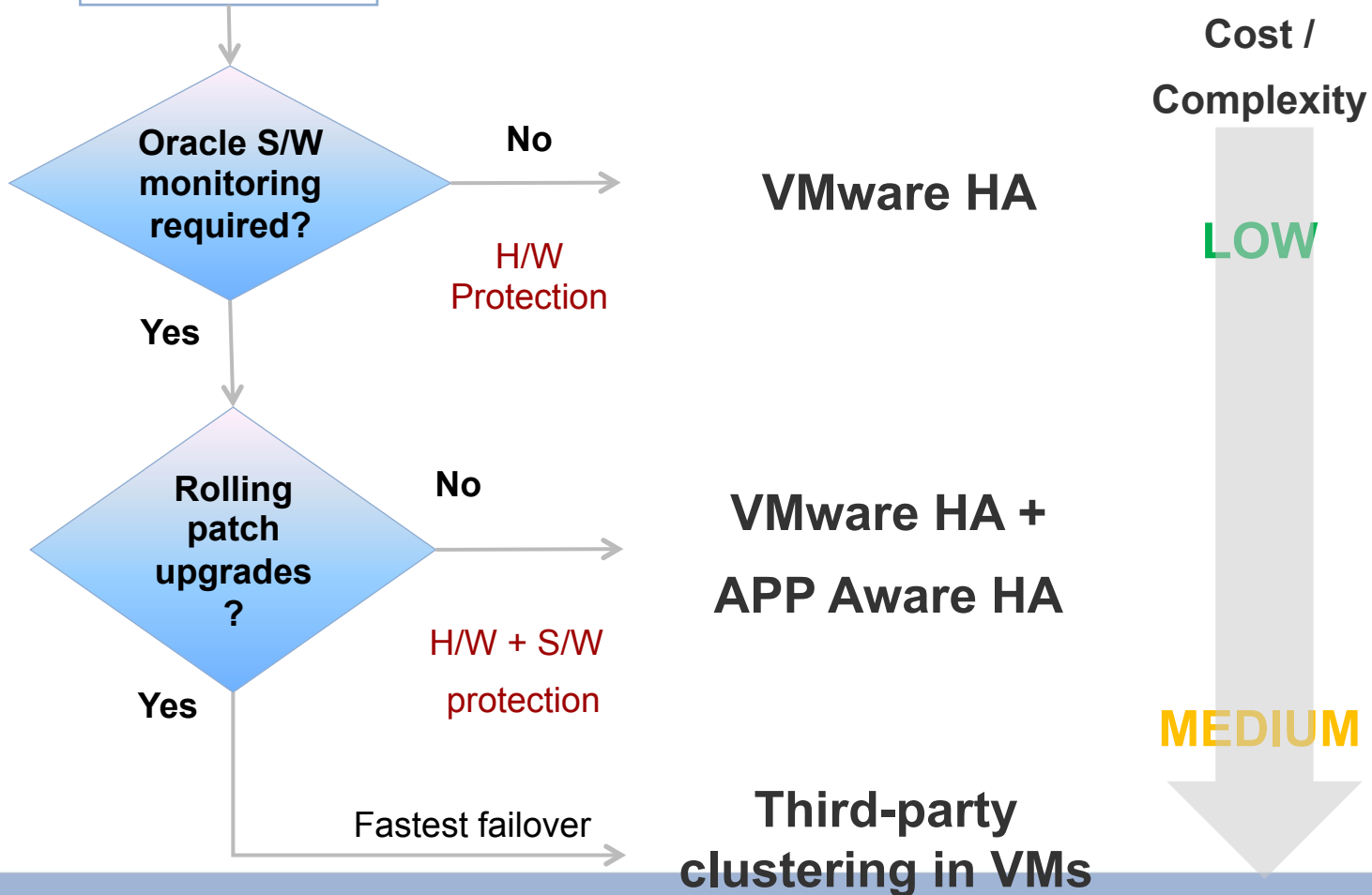
- ✓ VMware HA就是发生服务器故障是在其他的物理服务器上自动重启虚拟机

客户优势

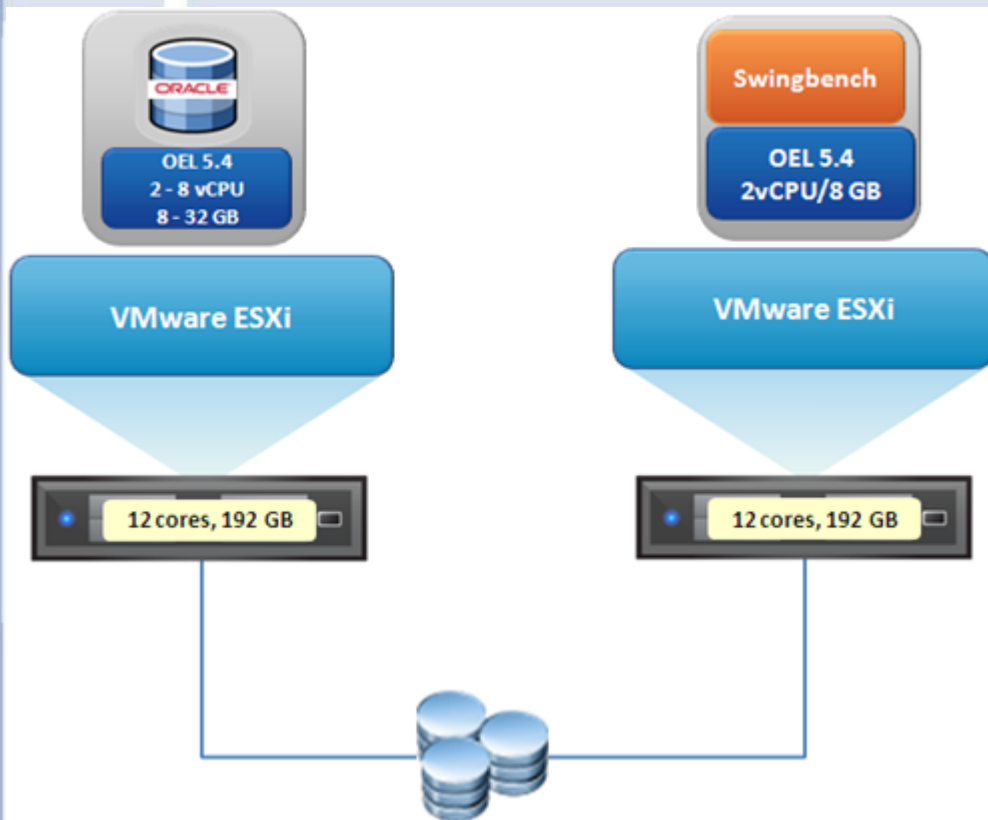
- ✓ 对所有的应用实现了高可用性，并且成本很低
- ✓ 实现硬件/OS级别的Monitor
- ✓ 不需要完全一致的重复硬件
- ✓ 比传统的集群有更高的成本优势，同时易于使用和操作

高可用方案的选择

由业务的 SLAs、
连续服务时间、
维护成本决定HA
方案



其他更高SLA等级的高可用设计(如Oracle RAC等)



典型场景：

- 云平台上部署多个数据库服务器，满足数据库服务器高可用需求

存在的问题：

- 云平台如何透过云平台的存储虚拟化功能实现底层Lun的共享
- 如何保障共享disk的前提下，仍能够使用热迁移、负载均衡的功能

HA总结

优势：

- 配置简单
- 能实现硬件、Hypervisor、OS、应用级别的高可靠

问题：

- 所有的业务数据、系统数据、应用存放在共享存储上，一旦共享存储发生故障，系统无法恢复

2种常见的的存储故障之一：人祸：VMFS分区表丢失

分区表丢失：一般是zone 或者 lun masking不当, 让物理主机和虚拟化主机的存储没有合理分区造成的

- 物理主机(比如Windows)将vmfs文件系统格式化

2种常见的的存储故障之一：天灾：物理存储硬件故障

物理存储发生故障、整个存储故障、链路故障或者LUN故障

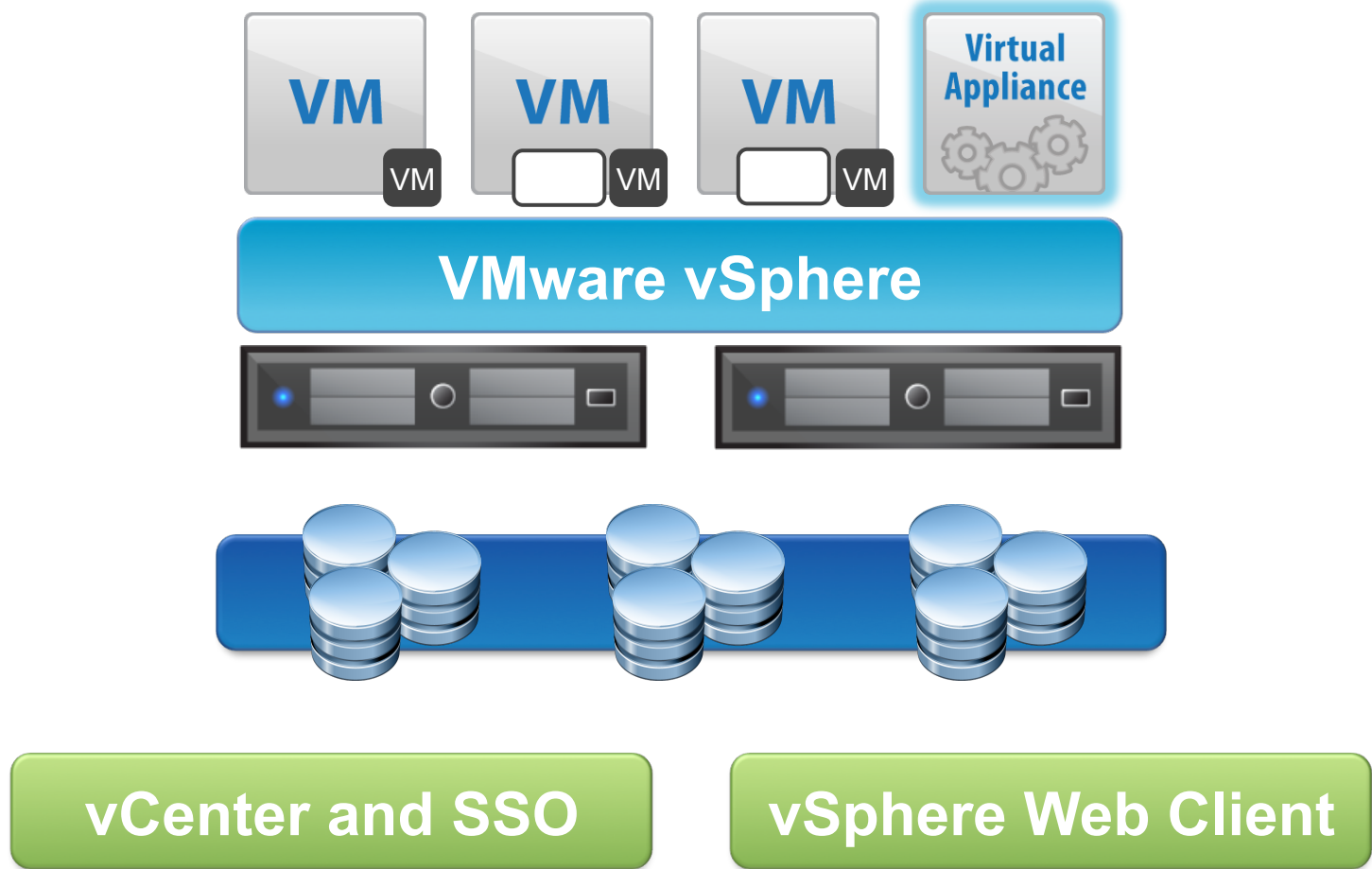
APD: all path down, 在5.0里面一般表示整个storage都发生了故障, 在5.0之前, APD会造成整个hostd无法连接, vcenter和主机失去连接, viclient无法连接, 5.0之后viclient可以连接, 但是速度缓慢

PDL: Permanent device lost, 在5.0里面一般表示一个lun发生故障, esxi主机从storage侧获得一个scsi探测信息, 会隔离错误, vcenter不会断开

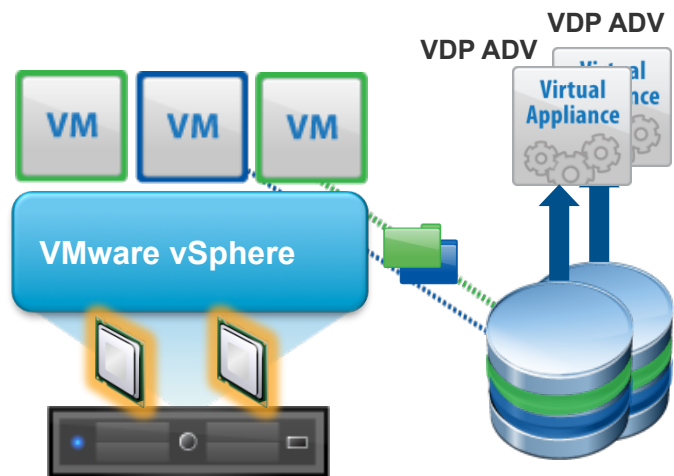
如何解决？

云资源池下的数据备份与恢复

VDPA 备份的架构



VDP Advanced



Overview

主要特点

- 支持重复数据删除、增量备份、全量备份、备份Schedule等
- 最大支持8T*10的重复数据删除后的容量
- 通过VSS支持应用的备份：SQL Server等
- 支持文件级别的恢复：各个VM可以自服务
- LAN FREE + SERVER LESS的备份

适用场景

- 用户数据的错误删除后的恢复，和快照相比有周期性等特点，且不会影响性能
- 对RTO要求不高的业务连续性要求
- 业务数据可以恢复到过去一个月甚至一年任意一个时间

Virtual Machine Images

Select this option if you want to backup virtual machine images.

Microsoft Applications

Select this option if you want to backup Microsoft Exchange and/or SQL Servers.

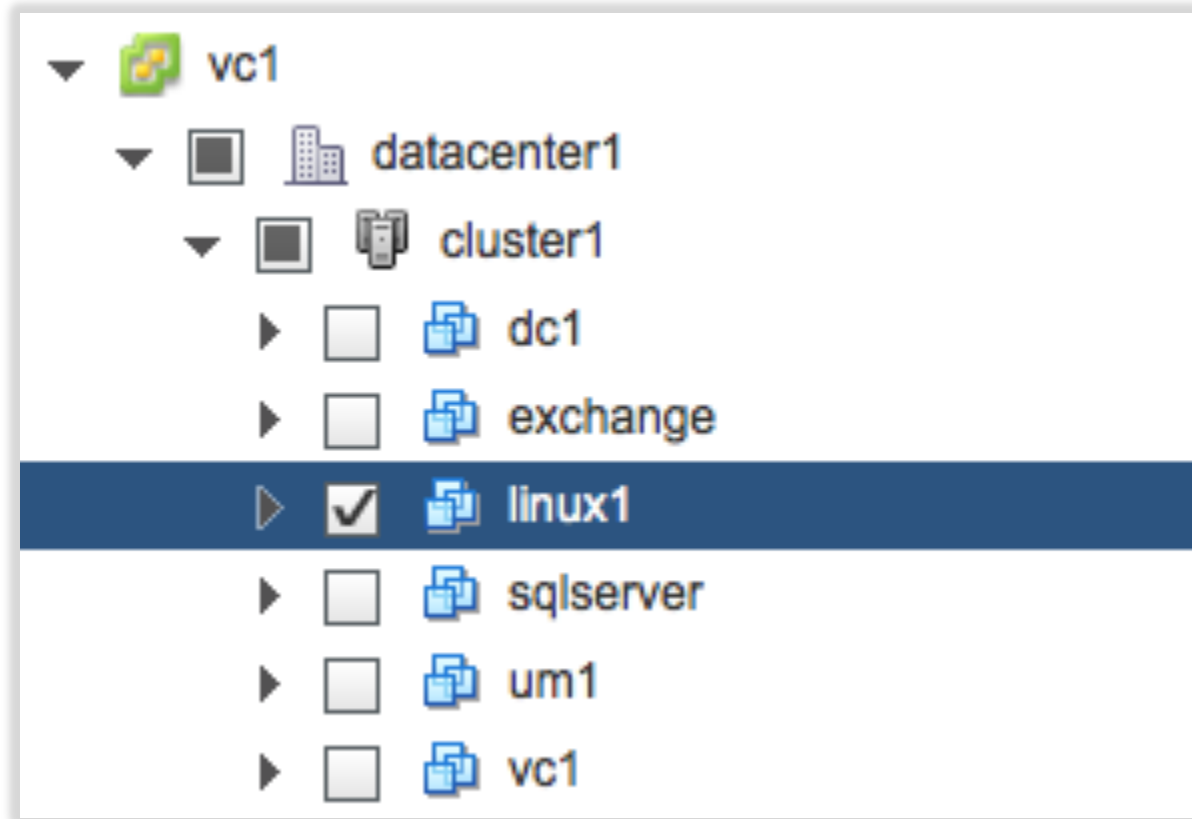
Full Server

Select this option to backup entire Exchange and/or SQL Servers.

Selected Databases

Select this option to backup selected databases from Exchange and/or SQL Servers.

Create Backup Job – Image



Create Backup Job – Image

Daily

Weekly performed every

The of every month

- Monday
- Tuesday
- Wednesday
- Thursday
- Friday
- Saturday**

备份周期

Forever

for

until

this Schedule:

Daily for:

Weekly for:

Monthly for:

Yearly for:

恢复整个虚拟机

Set the restore options for each backup that you are restoring.

Client: db-server

Backup: 08/04/2012 08:01 PM

Restore to Original Location

New Name: db-server_8_4_2012

Destination: /Datacenters/Datacenter Site A/Cluster A

Datastore: das2 (361.2 GiB free) ▼

文件级别的恢复(FLR)

最终用户进行自服务的文件恢复

通过WEB浏览器登录VDP Restore Client 进行文件恢复

- 用户无需安装备份和恢复使用的客户端
- <https://<VDP appliance IP address>:8543/flr>

Linux的用户和windows的用户都可以使用文件级别的恢复

File Level Restore (FLR)

The screenshot displays the 'Mounted Backups' window. On the left, a tree view shows the backup structure: 'Thu Jan 31 20:03:18 GMT-0500 2013' is expanded to show '1' drive, which contains several folders including '\$Recycle.Bin', 'PerfLogs', 'Program Files', 'Program Files (x86)', 'ProgramData', 'Recovery', 'System Volume Information', 'Users', and 'Windows'. The 'Users' folder is selected. On the right, a list of files and folders is shown with checkboxes for selection. The 'administrator.VMWARE' folder is checked, and the 'desktop.ini' file is highlighted.

Mounted Backups		Name	
▼	Thu Jan 31 20:03:18 GMT-0500 2013	<input type="checkbox"/>	Administrator
▼	1	<input checked="" type="checkbox"/>	administrator.VMWARE
▶	<input type="checkbox"/> \$Recycle.Bin	<input type="checkbox"/>	Default
▶	<input type="checkbox"/> PerfLogs	<input type="checkbox"/>	desktop.ini
▶	<input type="checkbox"/> Program Files	<input type="checkbox"/>	Public
▶	<input type="checkbox"/> Program Files (x86)		
▶	<input type="checkbox"/> ProgramData		
▶	<input type="checkbox"/> Recovery		
▶	<input type="checkbox"/> System Volume Information		
▶	<input checked="" type="checkbox"/> Users		
▶	<input type="checkbox"/> Windows		

DP总结

优势:

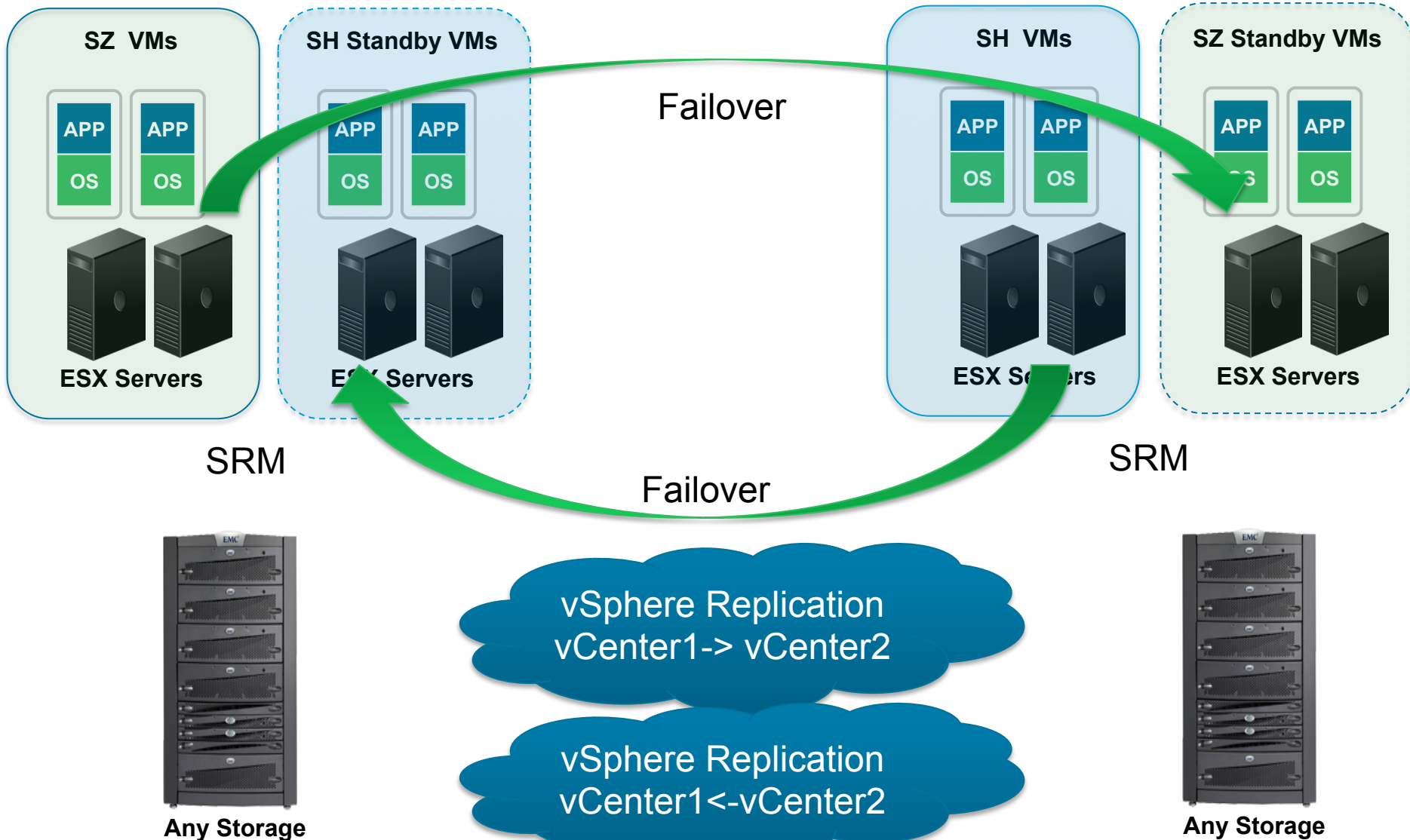
- 根据备份的策略, 能够恢复到过去一段时间内(一般为半年)内任意一个时间段内的虚拟机和文件

问题:

- 一旦发生大规模的存储硬件故障, VM的恢复速度受到存储写入速度的影响, 大大限制了业务的恢复时间
- 无法解决站点级别的故障

云资源池下的灾难恢复

利用SRM实现站点级业务连续性



SRM实现业务连续性

主要特点

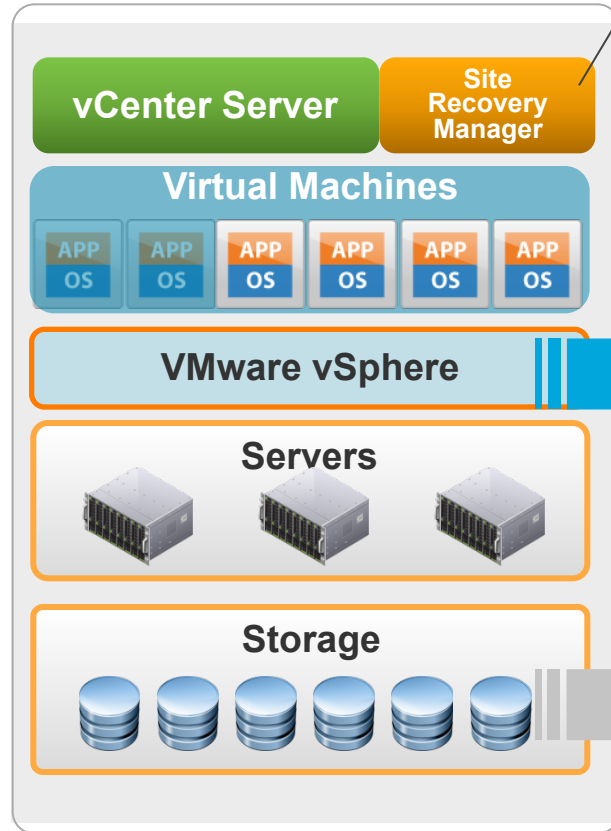
- 异构存储的数据复制: RPO=0~15分钟
- 快速的业务恢复: 500个虚拟机RTO=30分钟
- 数据的增量备份
- 虚拟机粒度的DR, 节省存储空间
- 快速可靠容灾演练
- 数据只保留一次最近的

适用场景

- 存储的硬件故障: 整个存储单点发生故障, 必须进行整个资源池的整体切换, 防止关键业务的长时间停机
- 站点的网络故障整体切换
- 站点的计算资源故障整体切换

SRM 5.x总体架构

Required at both protected and recovery sites



Site Recovery Manager

- 管理恢复计划、容灾演练
- 自动化failovers and failbacks
- 整合vcenter和复制计划

复制机制选择

vSphere Replication

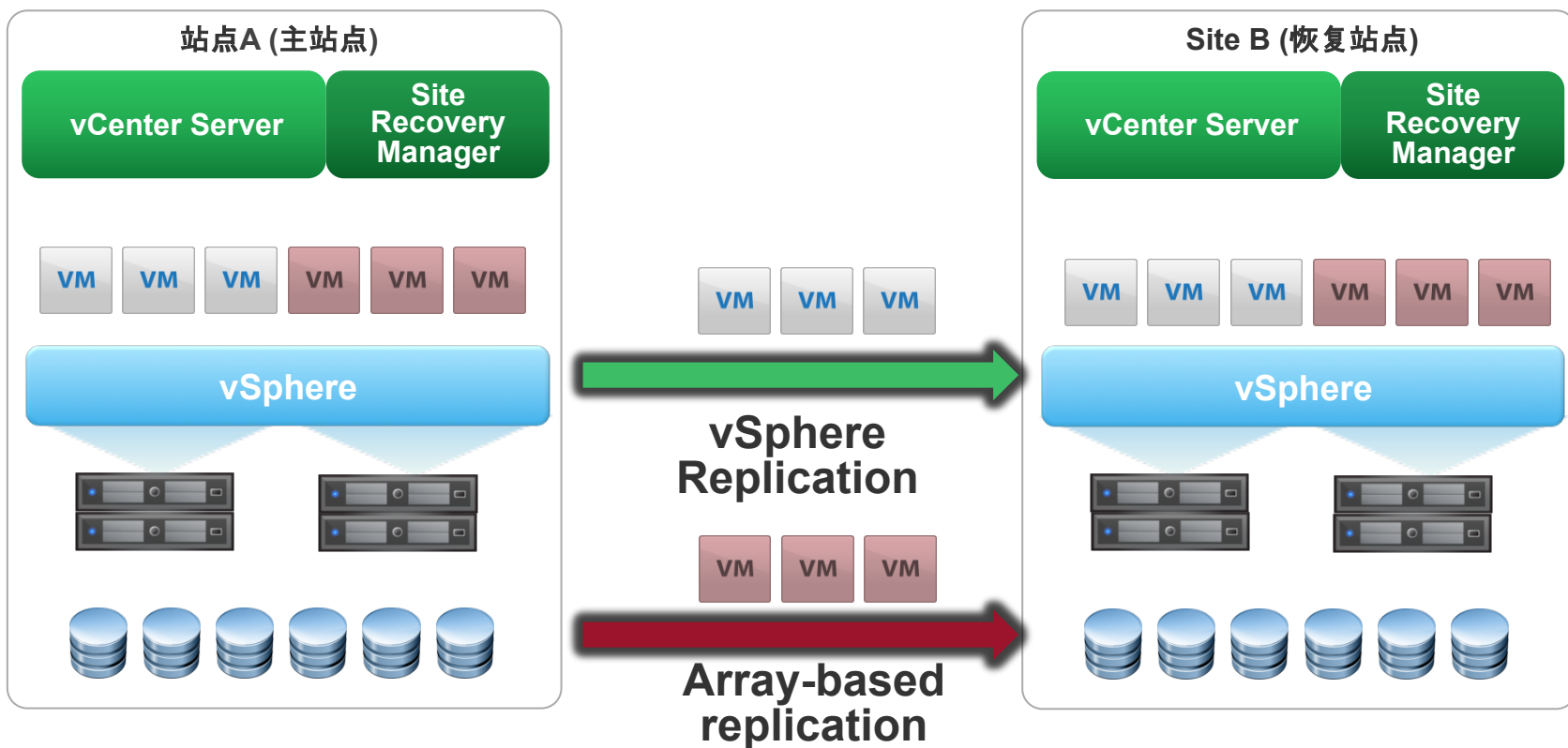
- Part of vSphere platform
- Replicates virtual machines between vSphere clusters

Storage-Based Replication (3rd party)

- Provided by replication vendor
- Integrated via replication adapters created, certified and supported by replication vendor



SRM 5.x 提供了复制选项的广泛选择



采用存储厂商的复制技术 (SRA: Storage Replication Agent)

The screenshot displays the SRM GUI interface. On the left, the 'Array Managers' pane shows a tree view with 'Site A - FS' selected. The main pane is titled 'Site A - FS' and has tabs for 'Summary', 'Array Pairs', 'Devices', and 'Permissions'. The 'Devices' tab is active, showing 'Devices for Enabled Array Pairs' and 'Devices for Array Pair: TS04-FGW-01 - TS04-FGW-02'. Below this, a table lists the devices and their replication details.

Device	Direction	Remote Device	Datastore	Protection Group
is_fgw1_p12_250_...	➔	TS04-FGW-01-is_fgw1_p12...	[is_fgw1_p12_250_01_repl]	PG 1 - App Collection X
is_fgw1_p12_250_...	➔	TS04-FGW-01-is_fgw1_p12...	[is_fgw1_p12_250_02_repl]	PG 2 - App collection Y

SRAs show detailed storage specific data in the SRM GUI: 基于LUN

採用vSphere Replication – VM 粒度

Track replication status through either web client or SRM

The screenshot shows the vSphere Web Client interface for vSphere Replication. The left sidebar shows a tree view with 'Sites' and 'PML-POD12-VC-A.pml.local'. The main content area is titled 'PML-POD12-VC-A.pml.local' and has tabs for 'Summary', 'Monitor', and 'Manage'. Under the 'Monitor' tab, there are sub-tabs for 'Issues', 'Outgoing Replications', and 'Incoming Replications'. The 'Incoming Replications' sub-tab is active, showing a table with the following data:

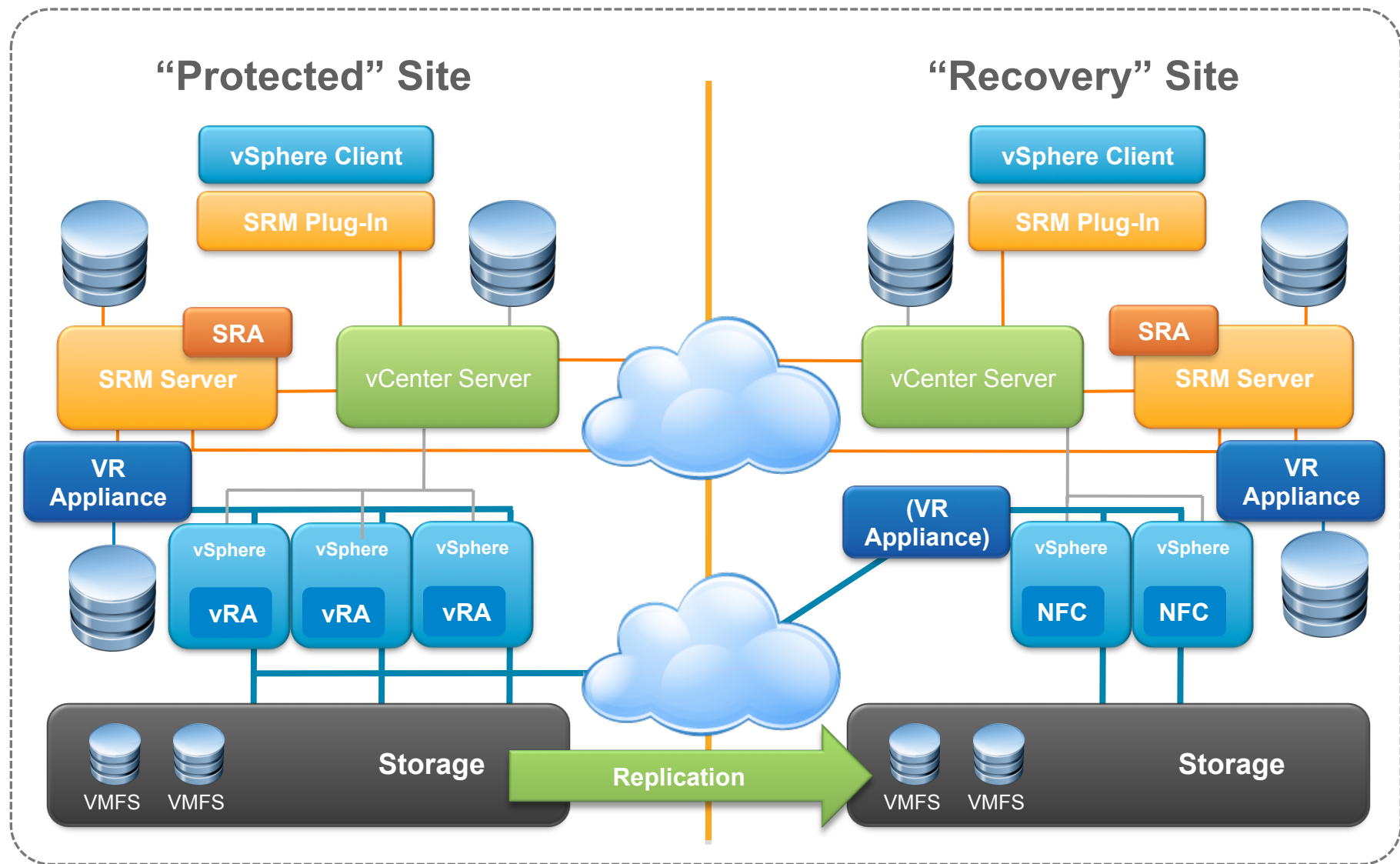
Virtual Machine	Status	Source	Completed	Duration	Size
VM-MultiDisk	OK	PML-POD12-VC-A.pm...	5/24/2012 7:21 AM	23 seconds	4.53 MB

The screenshot shows the vSphere Replication console for a replication job. The left sidebar shows a tree view with 'Sites' and 'tm-pod07-vrsb.tmsb.l...'. The main content area is titled 'tm-pod07-vrsb.tmsb.local' and has tabs for 'Summary', 'Virtual Machines', and 'Permissions'. Under the 'Virtual Machines' tab, there is a section for 'Source Site: Site A (Local)' and 'Target Site: Site B'. Below this, there are buttons for 'Move to...', 'Configure Replication...', 'Resume Replication', 'Pause Replication', and 'Remove Replication'. A table shows the replication progress for three VMs:

Virtual Machine	Replication Status	Last Sync Complet...	Last Sync Duratio...	Last Sync Size	RPO
VTestWK1	I.. 15%				04:00
VTestWK2	I.. 14%				04:00
VTestWK3	I.. 1%				04:00

Manage VR replication through SRM

SRM Architecture with vSphere Replication



■ VR redo log技术：

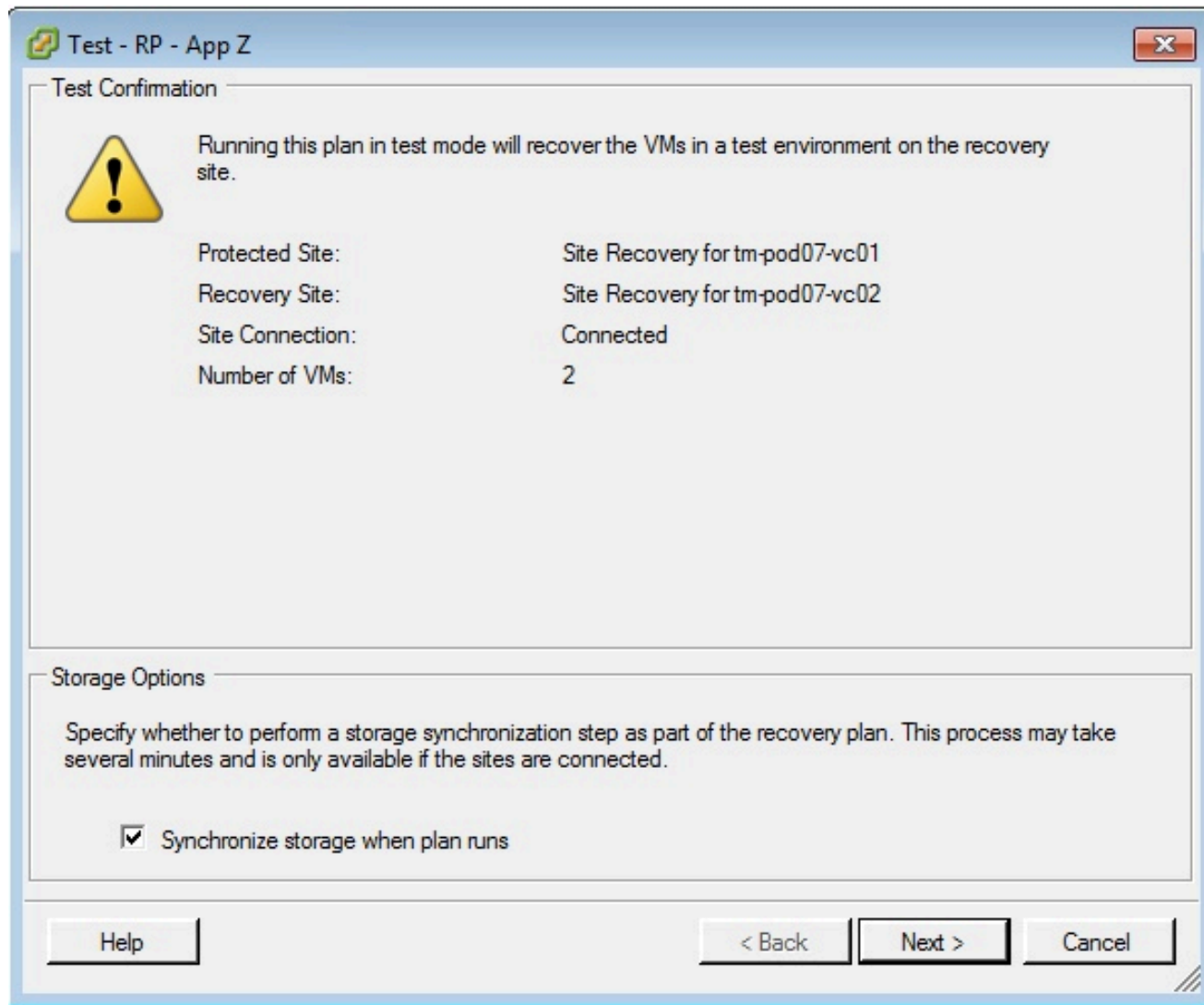
- VR通过一种轻量级的虚拟机快照，即redo log来记录上次RPO到本次RPO之间改变的数据块。假设虚拟机有多块虚拟磁盘，redo log可以保证多块虚拟磁盘之间的数据一致性。Redo log是通过虚拟SCSI过滤器来纪录数据块的读写情况，通过bitmap标记更改了的数据块，以便下次同步时传输。

DR技术信心的来源：容灾演练

Test

▶ Test

API



容灾演练 – 存储层/网络层

Protected Site



Recovery Site



Isolated Test Network



存储层: 增量的log继续保存, 现有的log合并后做快照



网络层: 接入网络设计一个隔离的vlan标签

容灾演练监控

RP - App Z Test In Progress

Edit Plan Export Steps Add Step Edit Step Delete Step Add Non-Critical VM

View: Test Steps

Recovery Step	Status	Step Started	Step Completed
1. Synchronize Storage	Success	4/28/2011 7:29:57 AM	4/28/2011 7:30:34 AM
1.1. Protection Group PG - FS - Application Group Z	Success	4/28/2011 7:29:57 AM	4/28/2011 7:30:34 AM
2. Restore hosts from standby			
3. Suspend Non-critical VMs at Recovery Site			
4. Create Writeable Storage Snapshot	Success	4/28/2011 7:30:34 AM	4/28/2011 7:31:17 AM
4.1. Protection Group PG - FS - Application Group Z	Success	4/28/2011 7:30:34 AM	4/28/2011 7:31:17 AM
5. Power On Priority 1 VMs			
6. Power On Priority 2 VMs			
7. Power On Priority 3 VMs	Running	4/28/2011 7:31:17 AM	50%
7.1. TestWK7	Running	4/28/2011 7:31:17 AM	66%
7.1.1. Configure Storage	Success	4/28/2011 7:31:17 AM	4/28/2011 7:31:22 AM
7.1.2. Power On	Success	4/28/2011 7:31:30 AM	4/28/2011 7:31:33 AM
7.1.3. Wait for VMware Tools	Running	4/28/2011 7:31:33 AM	0%
7.2. TestWK8	Running	4/28/2011 7:31:17 AM	35%
7.2.1. Configure Storage	Success	4/28/2011 7:31:17 AM	4/28/2011 7:31:26 AM
7.2.2. Power On	Running	4/28/2011 7:31:35 AM	6%
7.2.3. Wait for VMware Tools			
8. Power On Priority 4 VMs			
9. Power On Priority 5 VMs			

容灾演练完成：测试成功

Test Complete

The test has completed successfully. The virtual machines have been recovered in a test environment at the recovery site. Review the plan history to view any errors or warnings. When you are done with the test, press Cleanup to remove the test environment and reset this recovery plan to the ready state.

Recovery Step	Status	Step Started	Step Completed
1. Synchronize Storage	Success	4/28/2011 7:29:57 AM	4/28/2011 7:30:34 AM
1.1. Protection Group PG - FS - Application Group Z	Success	4/28/2011 7:29:57 AM	4/28/2011 7:30:34 AM
2. Restore hosts from standby			
3. Suspend Non-critical VMs at Recovery Site			
4. Create Writeable Storage Snapshot	Success	4/28/2011 7:30:34 AM	4/28/2011 7:31:17 AM
4.1. Protection Group PG - FS - Application Group Z	Success	4/28/2011 7:30:34 AM	4/28/2011 7:31:17 AM
5. Power On Priority 1 VMs			
6. Power On Priority 2 VMs			
7. Power On Priority 3 VMs	Success	4/28/2011 7:31:17 AM	4/28/2011 7:34:06 AM
7.1. TestWK7	Success	4/28/2011 7:31:17 AM	4/28/2011 7:34:06 AM
7.1.1. Configure Storage	Success	4/28/2011 7:31:17 AM	4/28/2011 7:31:22 AM
7.1.2. Power On	Success	4/28/2011 7:31:30 AM	4/28/2011 7:31:33 AM
7.1.3. Wait for VMware Tools	Success	4/28/2011 7:31:33 AM	4/28/2011 7:34:06 AM
7.2. TestWK8	Success	4/28/2011 7:31:17 AM	4/28/2011 7:33:36 AM
7.2.1. Configure Storage	Success	4/28/2011 7:31:17 AM	4/28/2011 7:31:26 AM
7.2.2. Power On	Success	4/28/2011 7:31:35 AM	4/28/2011 7:31:38 AM
7.2.3. Wait for VMware Tools	Success	4/28/2011 7:31:38 AM	4/28/2011 7:33:36 AM
8. Power On Priority 4 VMs			
9. Power On Priority 5 VMs			

VM's are ready to be used now

清楚Test Recovery



- 存储层:删除快照
- 网络层:删除临时的VLAN标签,恢复容灾站点的VLAN标签

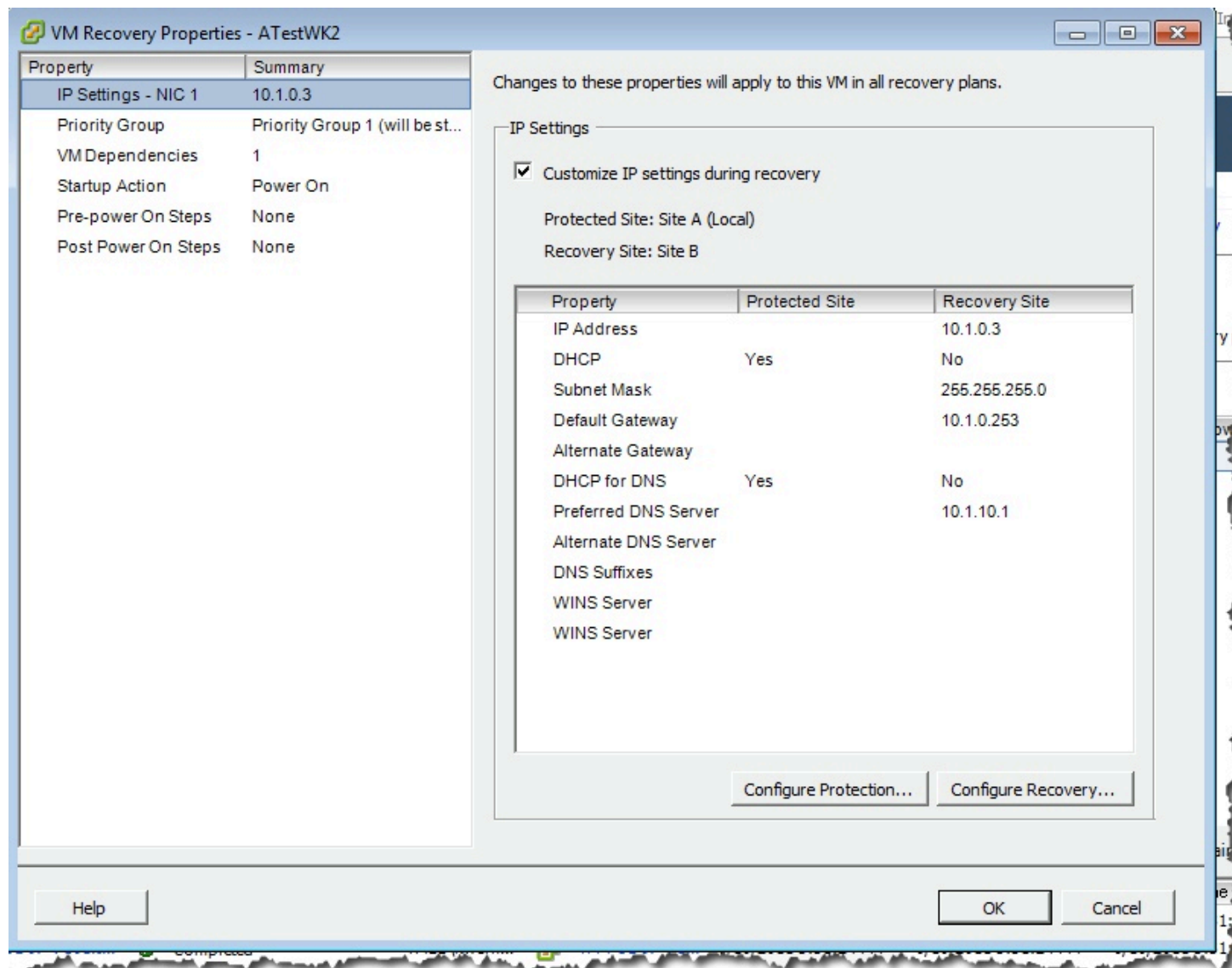
Disaster Recovery : 两种方式



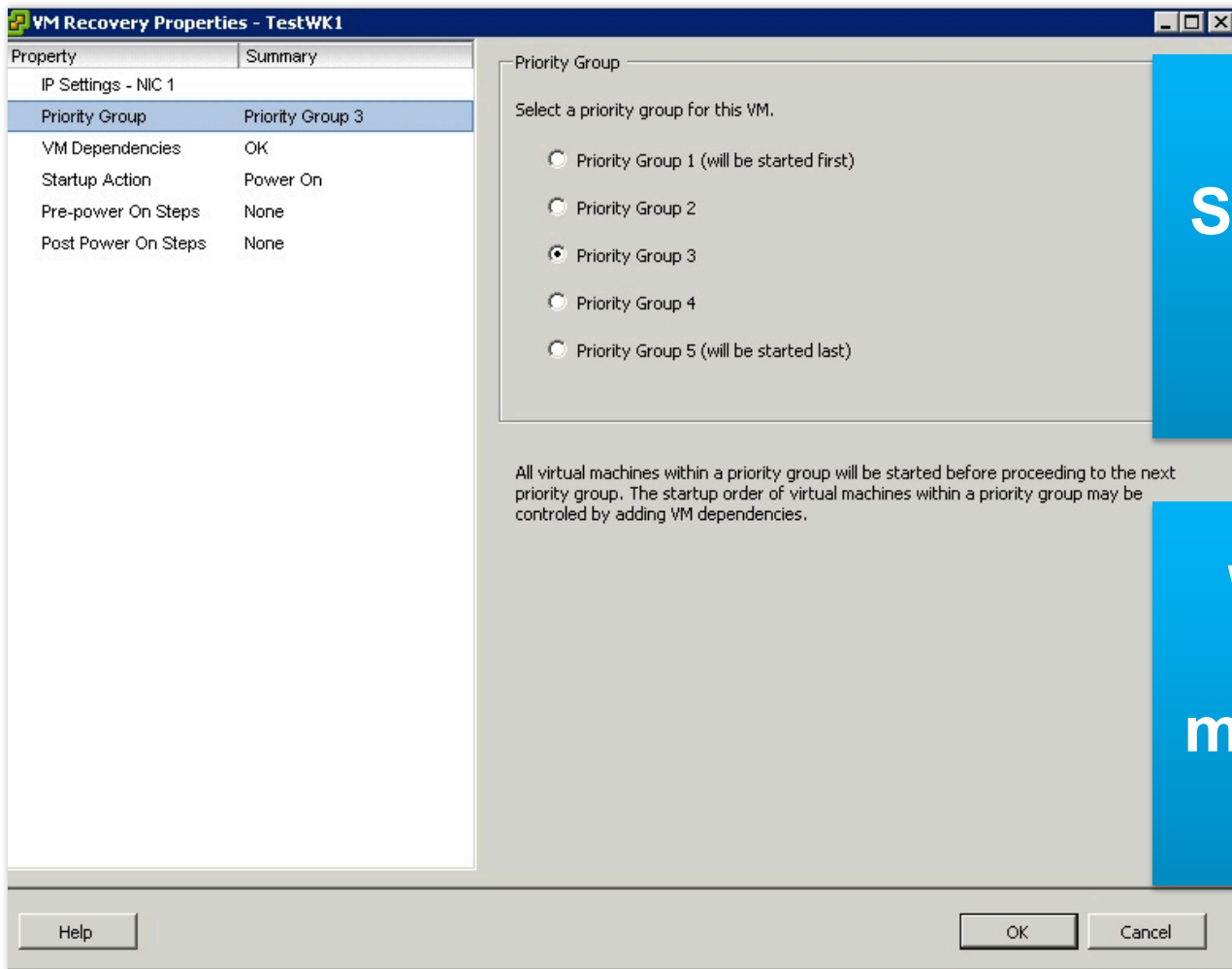
Will shutdown protected VM's, and than synchronize them IF it can!

Will NOT stop on errors and let you fix them!

IP 地址自定义



VM的关联性管理



SRM has 5 priority levels

Within a priority group all virtual machines will start simultaneously

RTO: 恢复时间

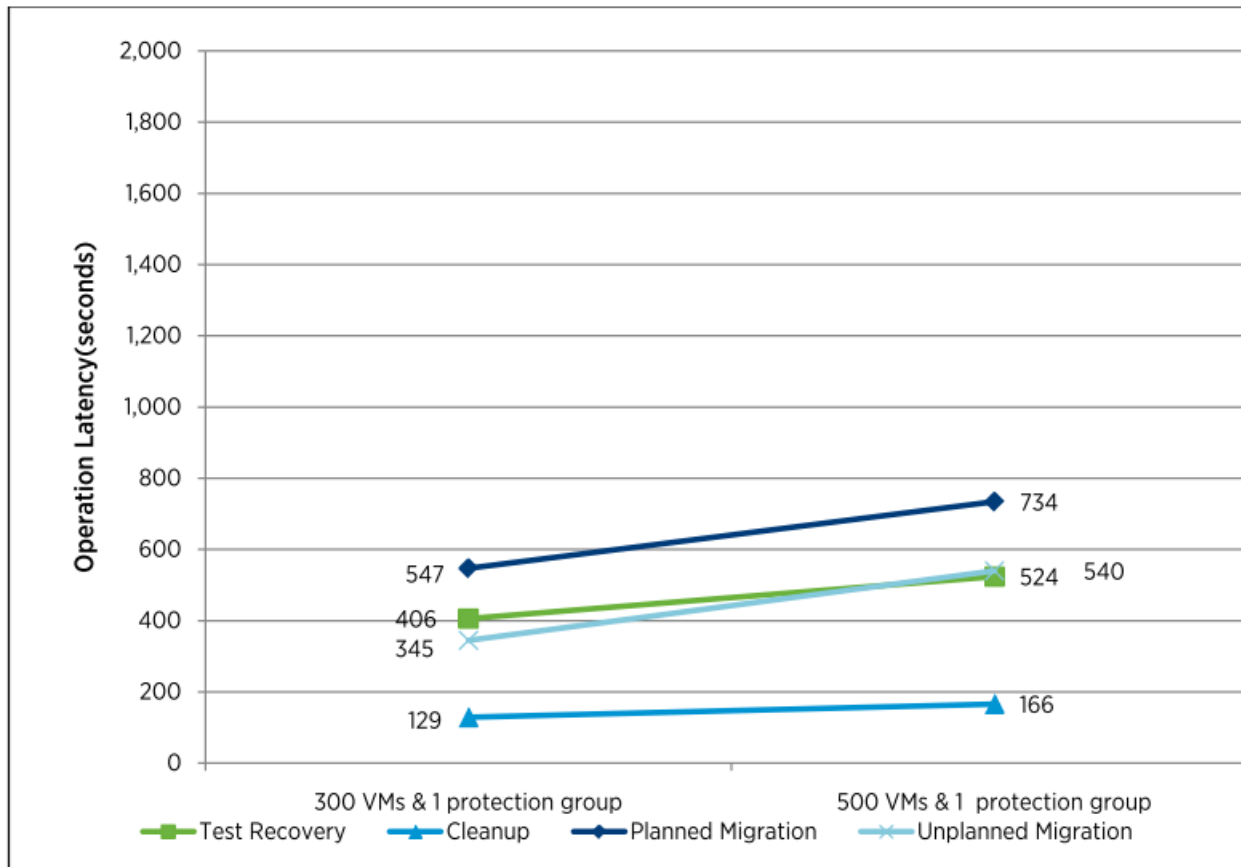


Figure 5. vSphere replication trending experiments: scaling with virtual machines

500个VM:

- 计划外容灾：
540s, 全部切换启动
- 计划内容灾：
734s, 全部切换

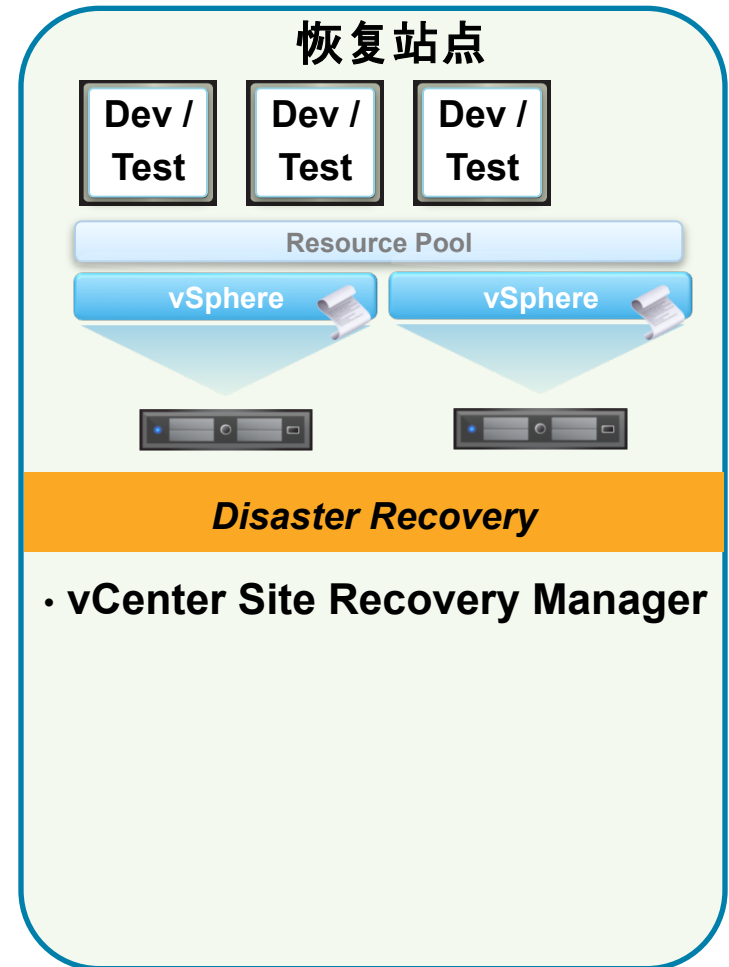
容灾是否一定需要大二层技术？LAN Extension

- SRM解决方案可以配置不用的业务系统VM在不同的站点使用的IP，所以和大二层技术没有因果关系
- 大二层技术必须配合路由优化技术使用
- 用户端访问数据中心可以通过域名的方式访问，可以通过三层方案实现DR
- 如果不使用域名访问业务，建议使用大二层+路由优化技术，或者发生容灾切换的时候修改业务访问地址

利用SRM实现数据中心“双活”方案特点

1. 实现两个数据中心同时提供服务，“双活”
2. 工作负载冷迁移，VM需要reboot
3. 无需两地同步存储投资低
4. 可以实现LAN Extension, 也可以做三层的方式，主要看用户的业务访问方式（IP/域名）
5. 对数据中心之间的网络环境要求较低,无延迟等要求
6. 在恢复站点可以实现定期的容灾演练

一个典型的VMware业务连续性解决方案



- Application and OS independent
- Focus on simplicity, cost-efficiency

其他双活技术:VMware Metro Storage Cluster 数据中心双活方案

方案描述

跨数据中心HA/DRS

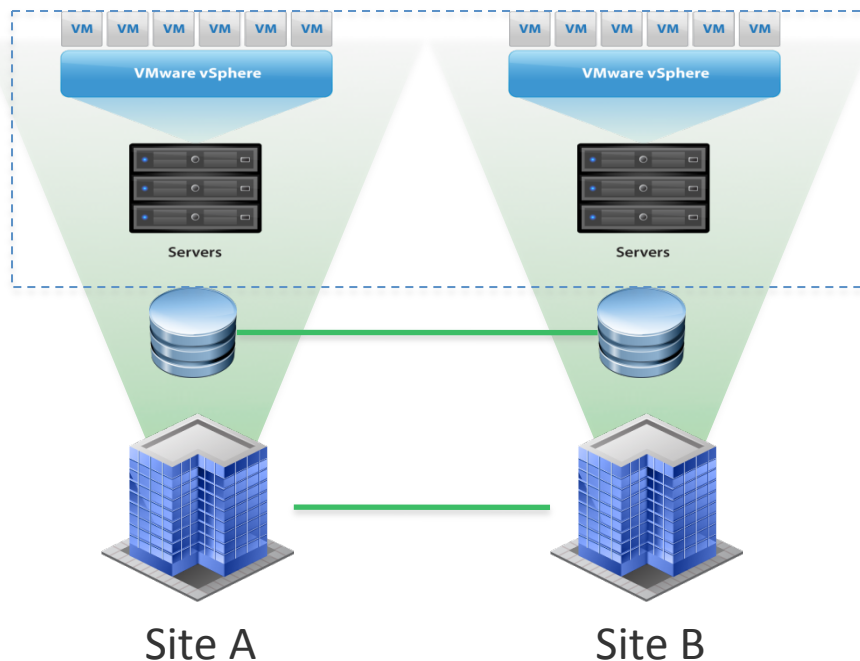
- 利用VMware HA, DRS Affinity 策略实现VM/主机绑定
- 需要分布式同步镜像存储
- 需要同一个二层网络
- 一个vCenter

用途

- 非计划性站点故障
- 存储故障
- 充分利用不同数据中心的计算资源

好处

- 实现跨数据中心工作负载的自动保护 (Disaster Avoidance)
- 高可用的效果如同本地集群(e.g. vMotion, DRS)
- 跨数据中心工作负载高可用和动态资源调整

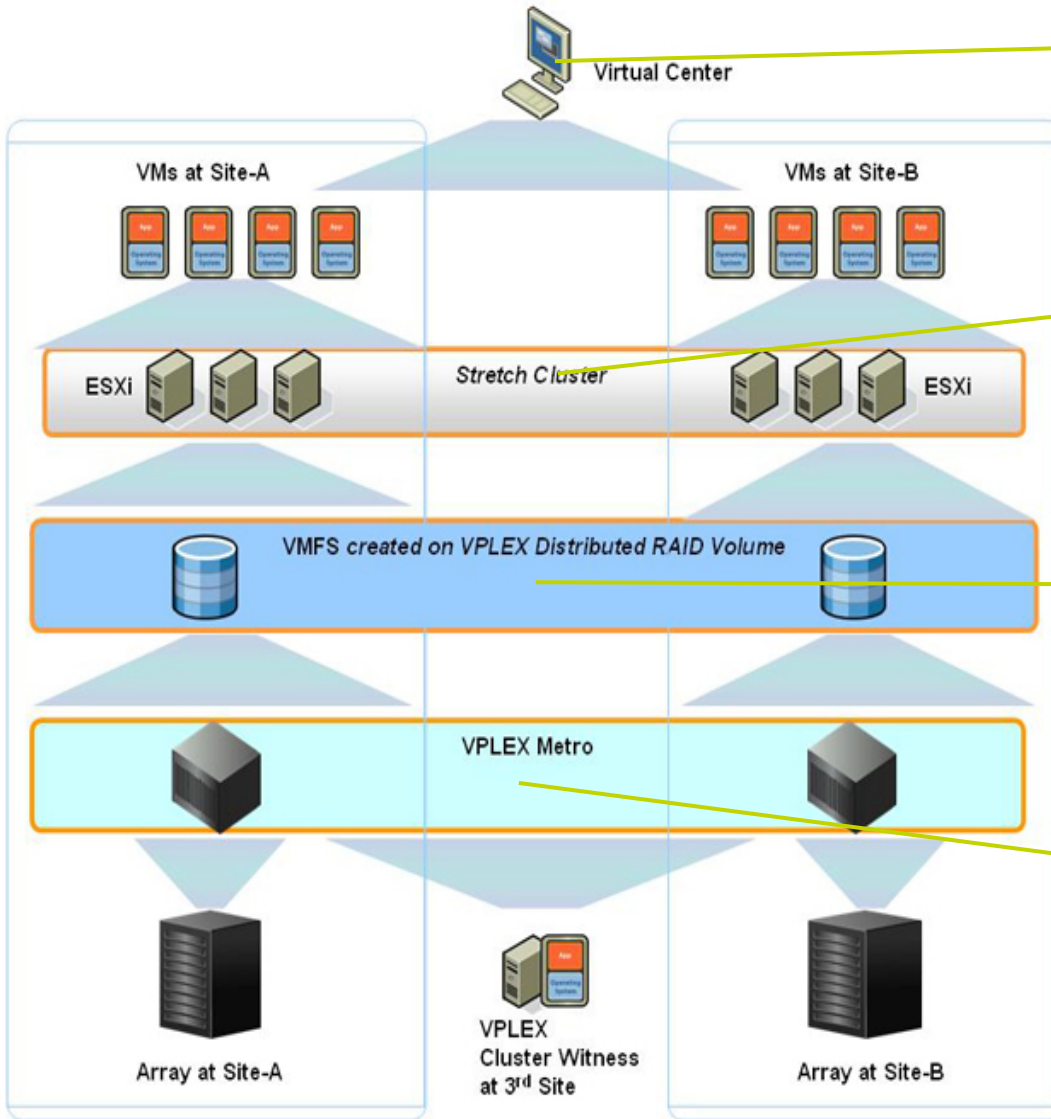


HA/DRS 集群

远程复制存储卷

二层网络扩展

vSphere Metro Storage Cluster 和长距离vMotion



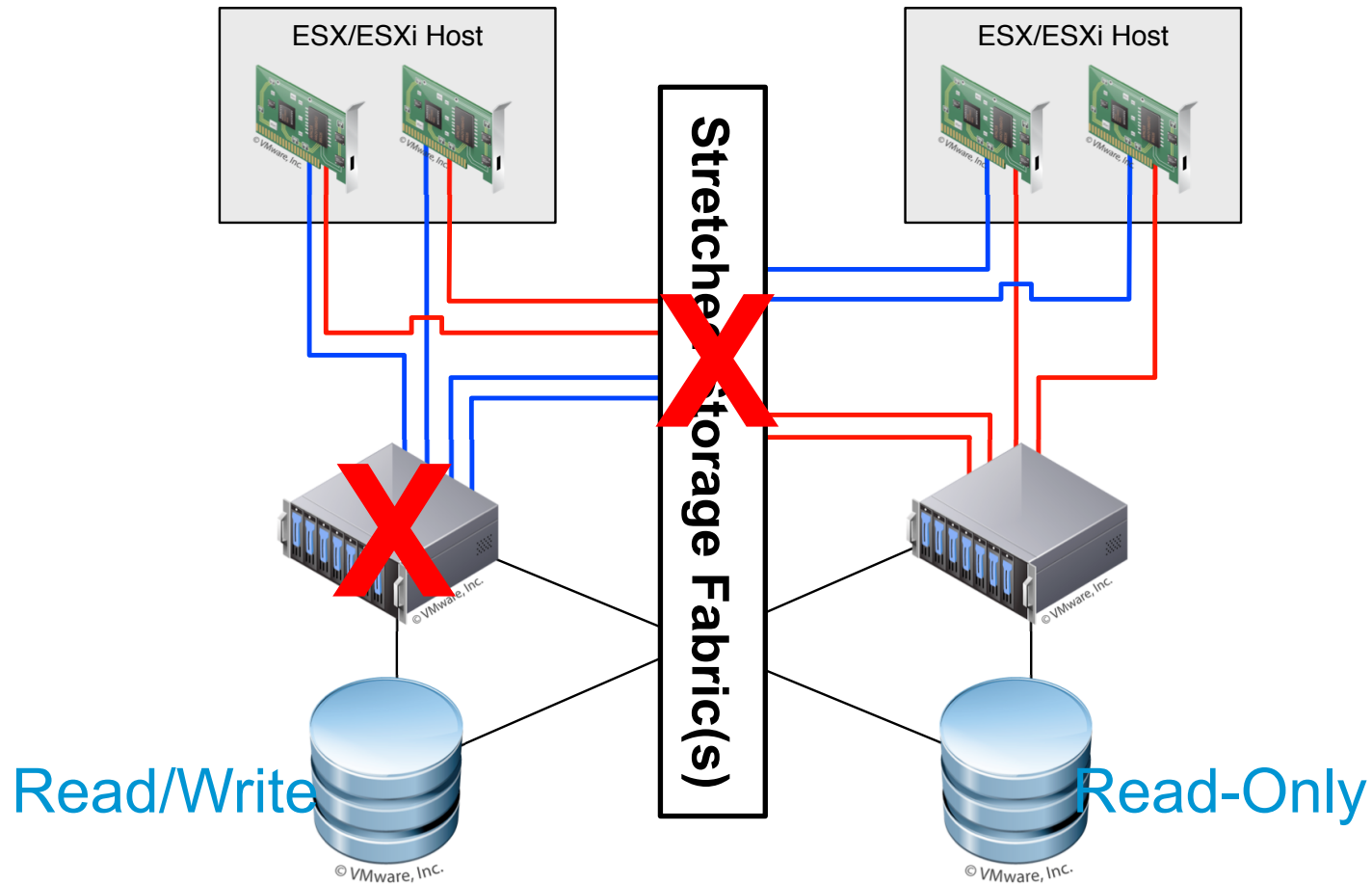
两数据中心ESXi主机由一套
vCenter管理
VC Heartbeat 做vCenter保护

两数据中心ESXi主机(vMotion, IP存储)
需要在同一个二层(IP)网段, <100KM

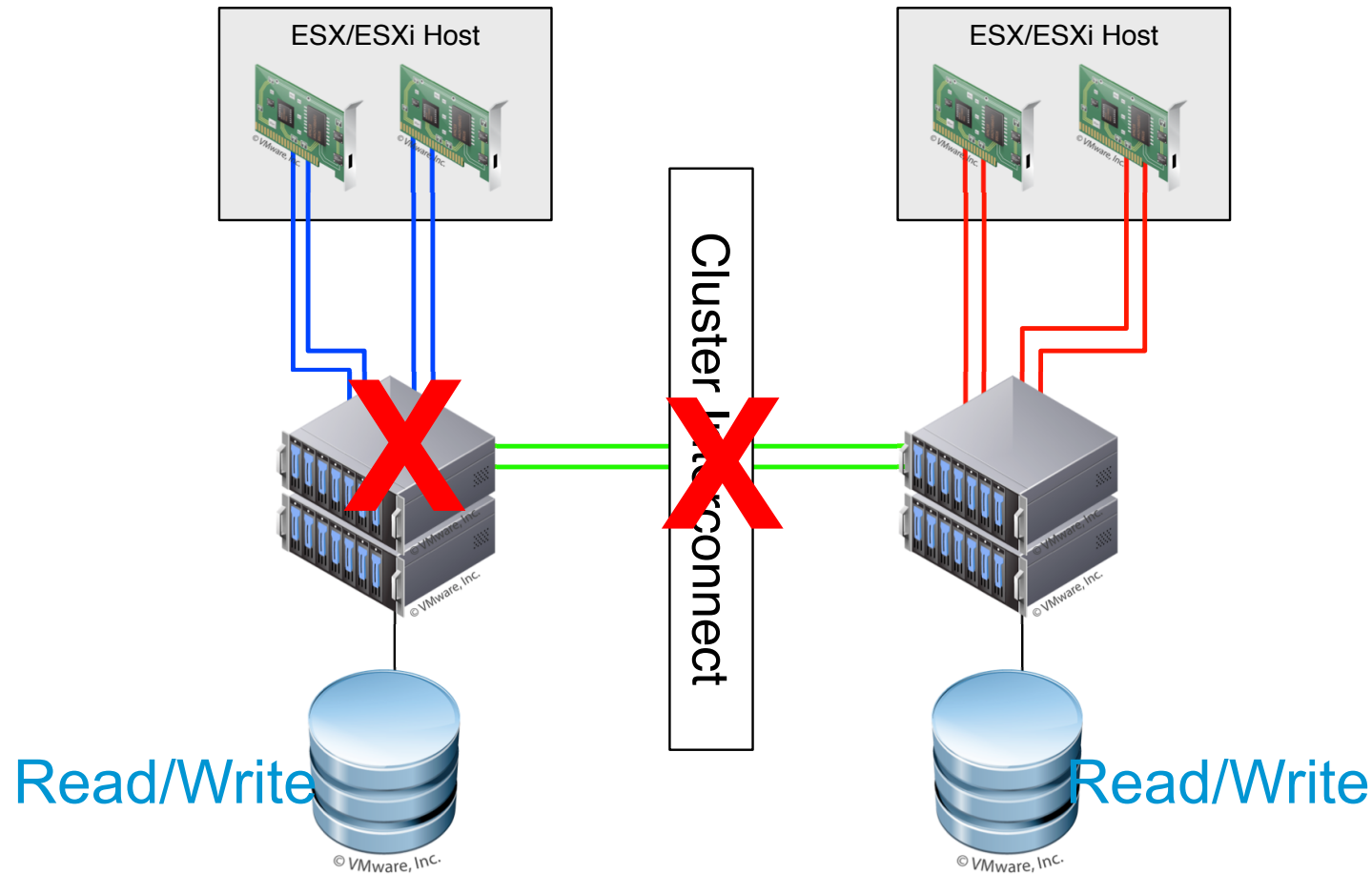
两数据中心之间须采用同步镜像存储,
支持FC 或者iSCSI

两数据中心之间存储网络互联,
支持FC 或者iSCSI, 一般通过是光纤
或者DWDM互联

“Uniform Access” Stretched Storage Model



“Non-Uniform Access” Stretched Storage Model



利用vSMC实现数据中心“双活”方案特点

1. 实现两个数据中心同时提供服务，“双活”
2. 工作负载热迁移，VM不需要停机
3. 两地同步存储投资高，目前支持6-7种存储
4. 必须实现LAN Extension
5. 对数据中心之间的存储网络环境和以太网网络环境要求较高（以太网要求RTT 延迟 $<10\text{ms}$ ；存储网各个厂商的设备要求不同，一般3-5ms的RTT延迟）
6. 数据中心广域网出入流量优化必须做！
7. 这种双活技术实际上是一种HA+DRS技术