云计算与虚拟化技术

第07章：Availability and Disaster Recovery

https://internet.hactcm.edu.cn

河南中医药大学信息技术学院互联网技术教学团队
河南中医药大学医疗健康信息工程技术研究所

2024.5

1

讨论提纲

- 高可用的分层模型
- vSphere HA
  - 配置实现
  - VMCP： Virtual Machine Component Protection
  - 主动式 HA： Proactive HA:
  - 准入控制： Admision control
- vSphere FT
  - 启用
  - FT-enabled VM 的操作
  - FT 对性能的影响
- Virtual Machine Clustering
- VMware Solutions

2

1. 高可用的分层模型

- **高可用（High Availability）是系统或服务在面对各种故障、异常情况或高负载时，仍然能够持续稳定地运行，并保持一定的服务能力。**
- **高可用的主要特点：**
  - 故障容忍性：
    - 能够承受部分组件的故障，如硬件故障、软件错误等，而不导致整个系统崩溃。
  - 快速恢复能力：
    - 当发生故障后，能够迅速恢复正常运行状态，减少服务中断的时间。
  - 持续服务：
    - 确保在大多数情况下用户都能持续地访问和使用服务。

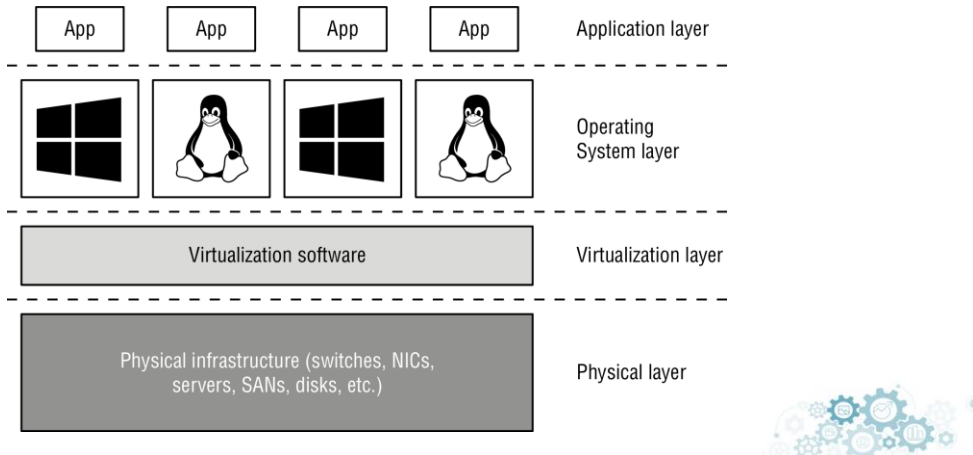
3

1. 高可用的分层模型

- **高可用（High Availability）是系统或服务在面对各种故障、异常情况或高负载时，仍然能够持续稳定地运行，并保持一定的服务能力。**
- **实现高可用的常用措施：**
  - 冗余设计：
    - 包括硬件冗余（如多服务器、备用电源等）和软件冗余（如备份数据、副本等）。
  - 负载均衡：
    - 将工作负载均匀分配到多个资源上，防止单点故障和过载。
  - 监控预警：
    - 实时监测系统状态，及时发现问题并发出警报。
  - 容错容灾：
    - 制定应对各种故障情况的策略和预案。

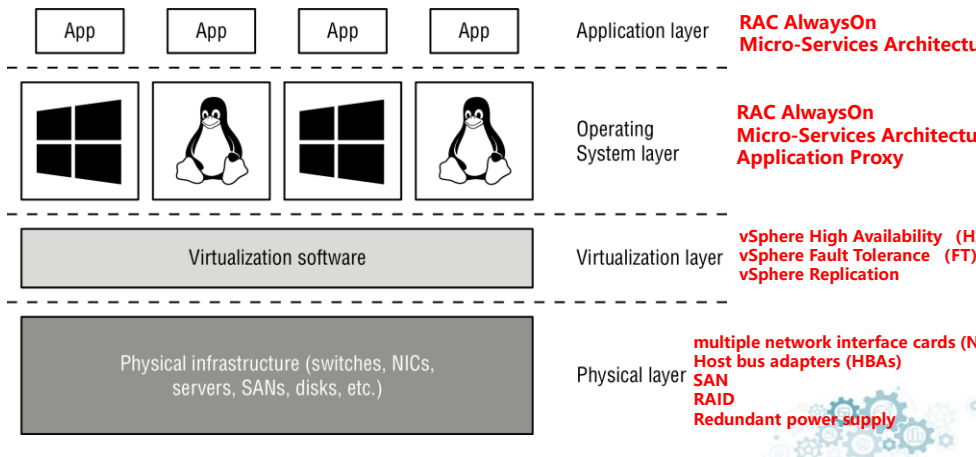
4

1. 高可用的分层模型



5

1. 高可用的分层模型



6

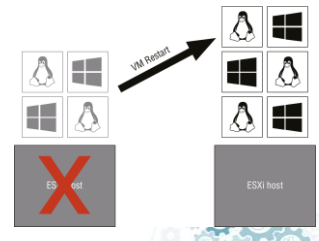
no one-size-fits-all solution
no a-package solution

没有“一刀切、一揽子”的解决方案

7

2. vSphere High Availability (HA)

- vSphere HA 主要用于确保虚拟机的高可用性。
  - 在群集中的多个主机上监测虚拟机的状态。
  - 当检测到某个主机出现故障时，自动在群集中的其他健康主机上重新启动受影响的虚拟机，从而最大程度地减少停机时间，保障业务的连续性。
- vSphere HA 保护级别：
  - 针对 ESXi Host 硬件故障的保护
  - 针对零停机计划内的维护
  - 针对 ESXi Host 计划外停机和灾难的保护

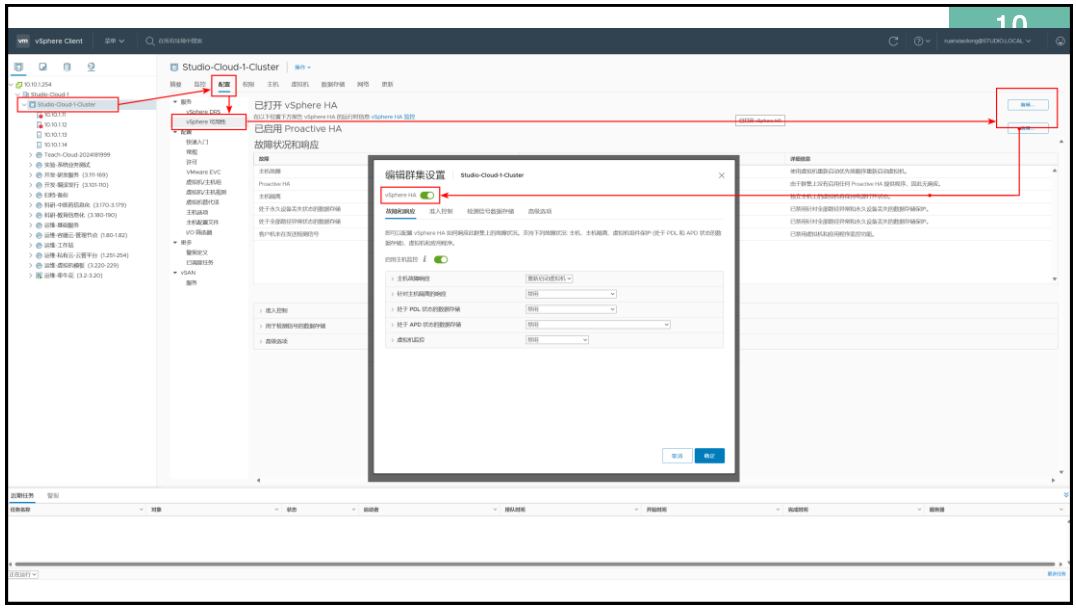


8

2. vSphere High Availability (HA)

- vSphere HA 启用的基本要求
  - vCenter Server
    - HA 必须依赖于 vCenter Server 才能实现，没有 vCenter Server 将无法启用HA。
  - 启用 vMotion
    - 当 ESXi Host 发生故障时，HA 会选择新的 ESXi Host 对虚拟机进行重新启动， 这个过程实质是迁移主机。
    - 迁移主机使用的技术是 vMotion，使用 HA 必须启用vMotion作为前提。
  - 网络冗余
    - HA 要求网络具有冗余功能，如无冗余的管理网络，HA 提示配置错误。
  - 安装 VMware Tools
    - 不仅添加了虚拟机驱动程序，部分 HA 检测机制也通过 VMware Tools 完成。
  - 群集 ESXi Host 数量要求：至少2个ESXi Host节点

9



10

2. vSphere High Availability (HA)

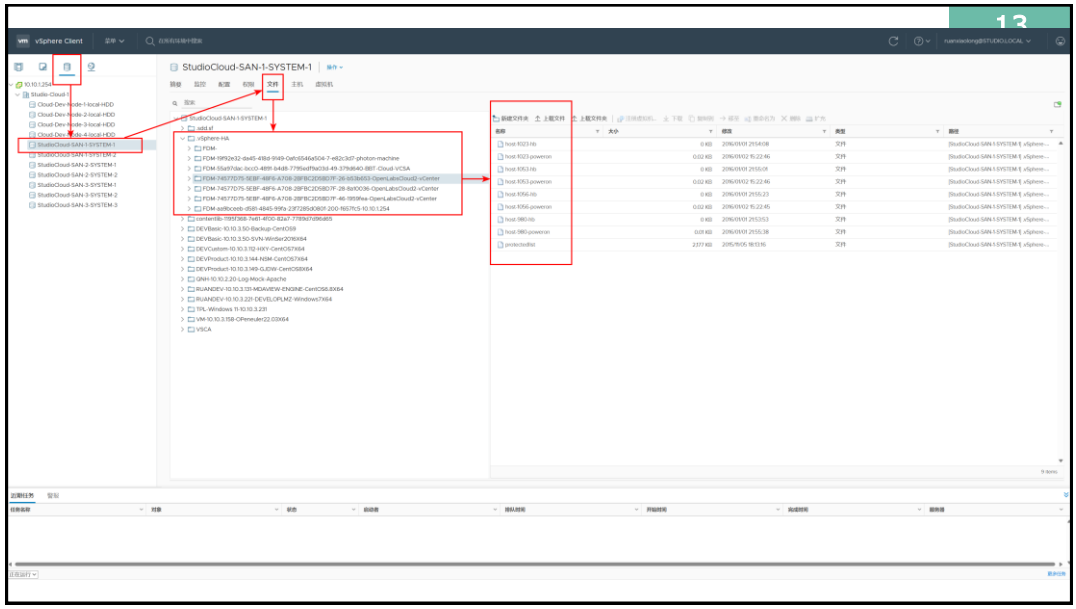
- vSphere HA 运行原理
  - 启用 HA时，系统会在群集中自动选举一台 ESXi Host 作为首选主机 (Master)，其余 ESXi Host 作为从属主机 (Slave) 。
  - Master 与 vCenter Server 进行通信，并监控所有受保护的从属主机状态。
  - Master 使用管理网络和数据存储检测信号来确定故障的类型。
    - 当 ESXi Host 故障时，Master 检测到并自动处理故障，让虚拟机重新启动。
    - 当 Master 本身出现故障时，Slave 会重新选举产生新的 Master。
- vSphere HA 群集中，主机故障由三种类型：
  - 故障：主机停止运行。
  - 隔离：主机出现网络隔离。
  - 分区：主机失去与首选主机的网络连接。

11

2. vSphere High Availability (HA)

- vSphere HA 如何工作?
  - 群集中每台 ESXi Host 均自行维护运行的 VM 列表。
  - VM 列表存储在共享存储的【vSphere-HA/<FDM集群ID>目录】。
  - 当心跳检测到 ESXi Host 主机故障时，则依据列表在群集中的其他 ESXi Host 启动 VM。

12

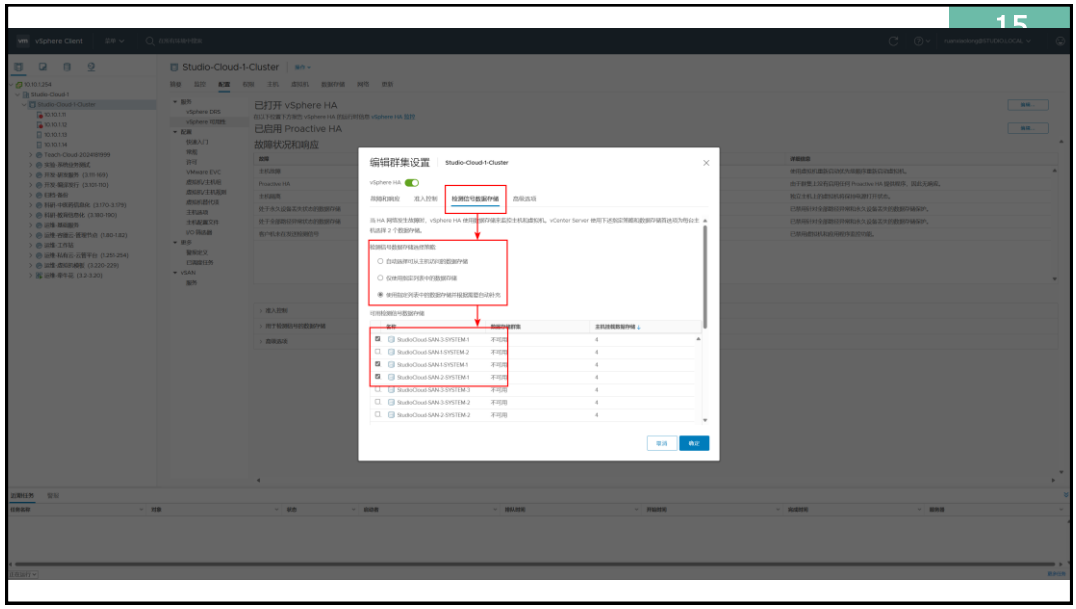


13

2. vSphere High Availability (HA)

- **心跳检测是用于确保 ESXi Host 已启动和运行。**
  - 如果在配置时间周期内没有接收到心跳，则认为 ESXi Host 已关闭。
  - 随后，触发 HA 事件处理机制。
    - 在正常运行 ESXi Host 上，重新启动无响应 ESXi Host 上的 VM。
- **vSphere HA 进行心跳检测的两种方式：**
  - Network heartbeat:
    - Master node 定期通过 ICMP 协议检查 Slave node 通过网络是否可以连通。
  - Storage heartbeat:
    - 群集中的每台 ESXi Host 都在共享存储上各存储一个空文件。
    - 共享存储上的文件由创建文件的 ESXi Host 锁定。
    - 当该文件未被锁定，则视为该文件对应的 ESXi Host 无法访问共享存储。

14

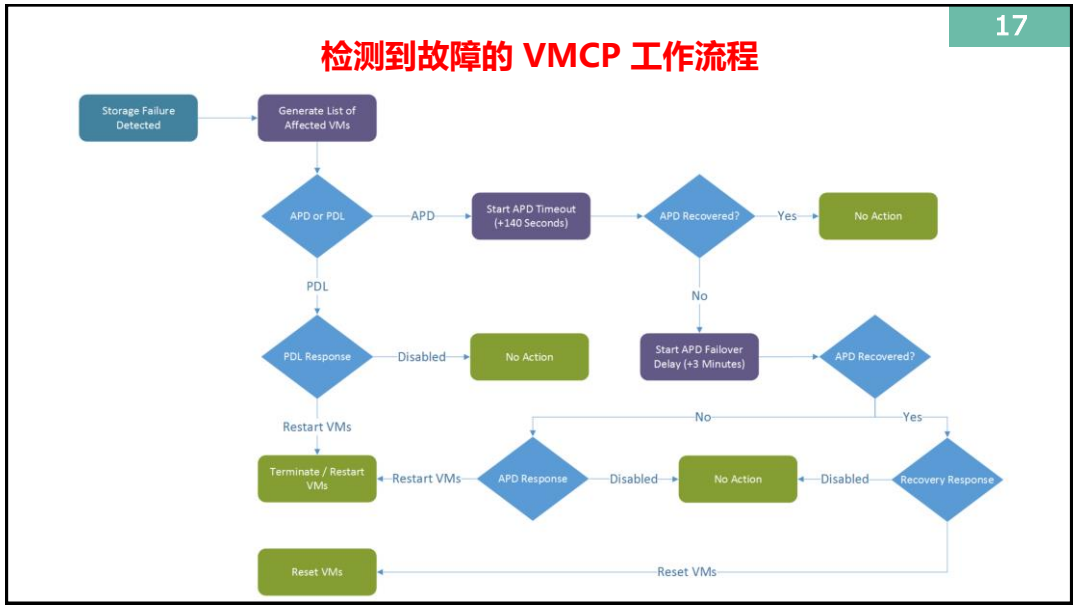


15

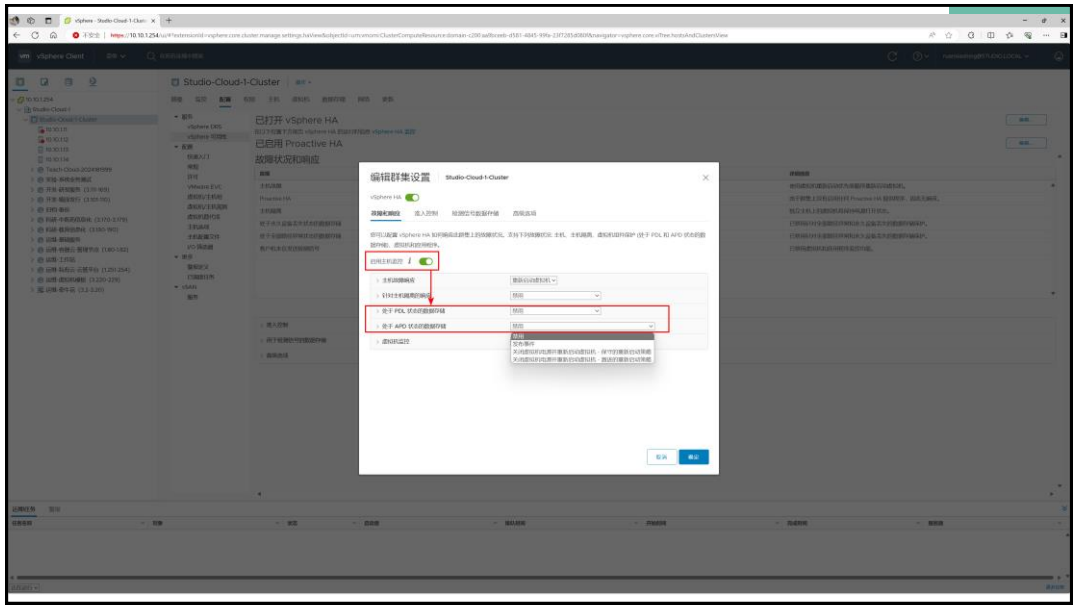
2. vSphere High Availability (HA)

- VMCP：Virtual Machine Component Protection,虚拟机存储保护
- 针对 VM，提供存储层面常见两种故障的支持。
  - PDL：Permanent Device Loss，永久设备丢失
    - PDL 是指存储设备出现永久性故障或以管理方式被移除或排除时的情况，预计该设备将来也不再可用。
    - 当存储阵列完全无法访问时，存储阵列将会发送一个 SCSI 指令到 ESXi Host。
    - ESXi Host 收到指令后，将停止向该存储设备进行 IO 操作。
  - APD：All Path Down，全部路径异常
    - APD 则是当存储设备的所有路径都出现故障，但没有迹象表明这是永久性还是暂时性设备丢失时的情况。
    - 如果 ESXi Host 无法访问存储阵列，且存储阵列也没有发送 SCSI 指令，则 ESXi Host 会将该存储设备标记为不可用。
    - 但是，ESXi Host 将不断尝试向存储设备发送 IO 指令，直到 APD 超时。

16


17
检测到故障的 VMCP 工作流程
Storage Failure Detected
Generate List of Affected VMs
APD or PDL
APD
Start APD Timeout (+140 Seconds)
APD Recovered?
Yes
No Action
PDL
No
PDL Response
Disabled
No Action
Start APD Failover Delay (+3 Minutes)
APD Recovered?
Restart VMs
No
Yes
Terminate / Restart VMs
Restart VMs
APD Response
Disabled
No Action
Disabled
Recovery Response
Reset VMs
Reset VMs

17

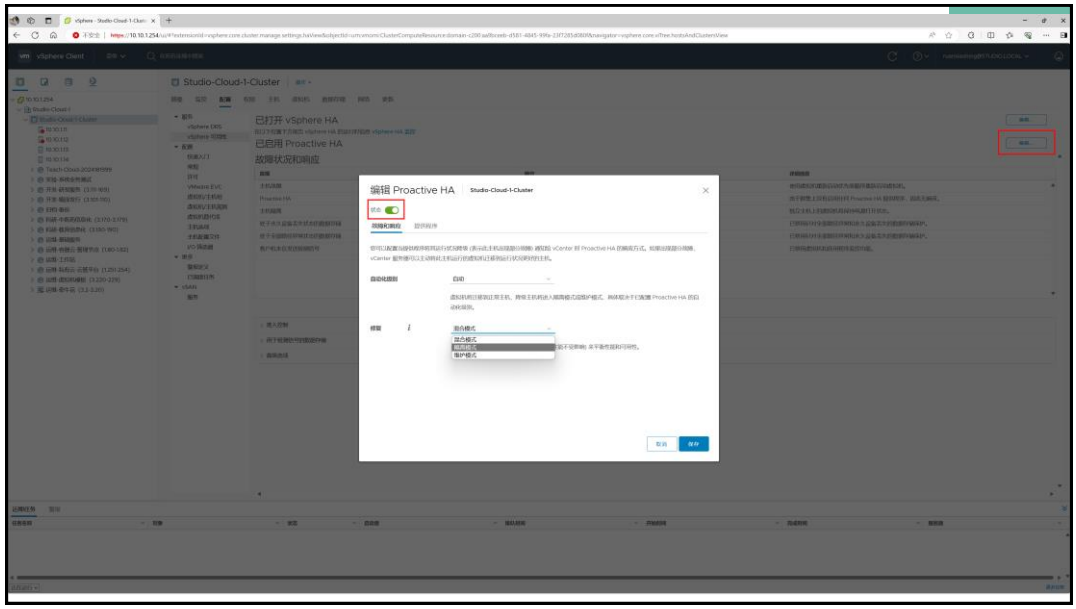

Studio-Cloud-1-Cluster
已打开 vSphere HA
已启用 Proactive HA
故障状况和响应
编辑群集设置
Studio-Cloud-1-Cluster

18

2. vSphere High Availability (HA)

- Proactive HA：主动预防系统故障
  - 传统的 HA 是被动的，只有当服务器故障时，才会把受保护的 VM 转移到其他的服务器上去。现在主流服务器厂商提供硬件系统监控和预警功能，使主动预防成为可能。
  - vSphere 6.5及以后版本，可以通过接口与以下服务器厂商的系统管理工具相集成，以实现主动预防式 HA (Proactive HA)，例如：Dell Openmanage、HP Insight Manager、Cisco UCS Manager
  - 服务器厂商的系统管理工具会把服务器的一些异常状况向 vSphere 告警。
    - 例如某个散热风扇发生故障，某块硬盘的读写故障率超出正常阈值范围等等。
    - 当告警发生时，意味着服务器处于亚健康状态，vSphere 就会把这台服务器处于隔离模式（Quarantine mode），这意味着该服务器上不会再启动新的虚机，并且 vSphere 会尽可能地把该服务器上的虚拟机 vMotion 到其他服务器上去。

19

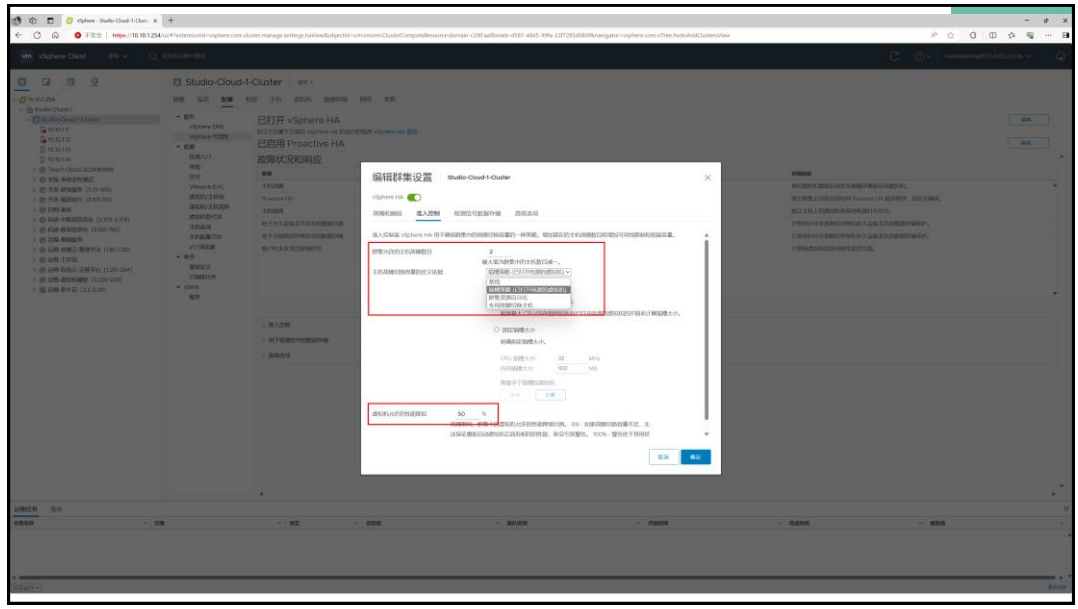


20

2. vSphere High Availability (HA)

- Admission control：准入控制
  - vSphere HA 使用准入控制确保在主机出现故障时预留足够资源用于恢复。
  - vSphere HA 准入控制的基础是群集保证可故障切换的主机故障数。
  - 可通过三种方式来设置主机故障切换容量：
    - 群集资源百分比
    - 插槽策略
    - 专用故障切换主机
  - 准入控制对资源使用施加限制，任何可能违反限制的操作都不会被允许。
    - 启用准入控制机制后，可能不允许的操作：
      - 打开虚拟机电源
      - 迁移虚拟机
      - 增加虚拟机的 CPU 或内存预留

21



22

2. vSphere High Availability (HA)

2.6 Admission control

vSphere HA 准入控制：
https://docs.vmware.com/cn/VMware-vSphere/6.7/com.vmware.vsphere.avail.doc/GUID-53F6938C-96E5-4F67-9A6E-479F5A894571.html

3. vSphere Fault Tolerance (FT)

3.1 了解 FT

- vSphere Fault Tolerance (FT) 是提高关键 VMs 可用性的技术，旨在实现 VMs 零宕机 (a zero-downtime technology)。
- vSphere FT 的工作原理：
  - 在两台 ESXi Host 上连续复制同一个VM的状态：
    - the primary VM，主VM
    - the secondary VM，辅助VM
  - 由于辅助虚拟机与主虚拟机的执行方式相同，并且辅助虚拟机可以无中断地接管任何状态下的执行，因此可以提供容错保护。
    - 如果运行主虚拟机的 ESXi Host 发生故障，将会执行透明故障切换，立即启用辅助虚拟机以替换主虚拟机。同时创建新的辅助虚拟机，自动重新建立 FT 冗余。
    - 如果运行辅助虚拟机的 ESXi Host 发生故障，在立即在新的 ESXi Host 上创建故障主机上的辅助虚拟机，实现 FT 冗余。

3. vSphere Fault Tolerance (FT)

- vSphere FT 的功能限制：
  - 每个 VM 只能够具有最多 8 vCPU。
  - 每台 ESXi Host 上最多支持 4 个辅助虚拟机。
  - 使用 FT 的 ESXi Host 必须使用相同系列的 CPU、配置 EVC。
  - 使用 FT 的 ESXi Host 必须使用 10Gbps 的网络，推荐使用专网。
  - 启用 FT 的虚拟机不支持下面的 vSphere 高级功能：
    - Storage vMotion
    - Linked clones
    - VMCP
    - Virtual volume datastores
    - Storage-based policy management
    - Snapshots

25

Comparing FT and SMP-FT

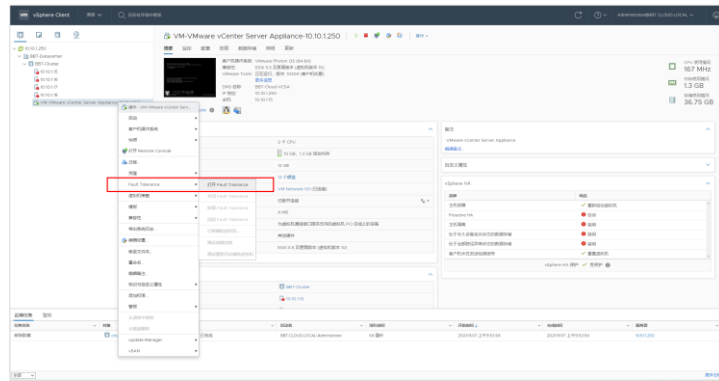
	FAULT TOLERANCE	SMP FAULT TOLERANCE
# CPUs supported	1	≤8
Memory virtualization hardware assist	Not supported	Supported
Disk format	Eager zero thick	Thin provisioning
VMDK redundancy	Not supported	Mandatory
VADP backups	Not supported	Supported
Required network bandwidth	1 GB	10 GB
DRS	Partially supported	Partially supported
Protected VMs per host	≤4	≤4
Paravirtualized Devices	Not supported	Supported

26

3. vSphere Fault Tolerance (FT)

3.2 实现 FT

- vSphere FT 的启用非常简单，但是必须基于单台 VM 进行配置。



27

3. vSphere Fault Tolerance (FT)

3.3 FT-enabled VM 的操作

- VM 启用 FT 后，增加的与 FT 相关的操作有：
  - 关闭：Turn Off Fault Tolerance
    - 将删除辅助虚拟机及其配置以及所有历史记录。
  - 恢复：Suspend Fault Tolerance
  - 挂起：Resume Fault Tolerance
  - 迁移辅助虚拟机：Migrate Secondary
    - 在为主要虚拟机打开 vSphere Fault Tolerance 之后，可以迁移其关联的辅助虚拟机。
  - 测试故障恢复：Test Failover
    - 可以通过诱发所选主要虚拟机的故障切换来测试容错保护。
  - 测试重新启动辅助虚拟机：Test Restart Secondary
    - 可以通过诱发辅助虚拟机发生故障以测试为所选主要虚拟机提供的容错保护。

28

3. vSphere Fault Tolerance (FT)

- FT 使虚拟机和应用程序的停机时间为零，但虚拟机的性能会受到影响。
  - 受 FT 保护的 VM 的性能会降低。
    - 每个 CPU 指令、内存更改或存储 I/O 都需要复制到辅助 VM。
  - 根据工作负载，FT 还可能生成大量网络流量。

29

Kernel Compile

This experiment shows the time taken to do a parallel compile of the Linux kernel. This is a both a CPU- and MMU-intensive workload due to the forking of many parallel processes. During this benchmark the CPU was 100 percent utilized. This workload did some disk reads and writes, but generated no network traffic. As seen in Figure 2, FT protection increases the kernel compile time a small amount—about 7 seconds.

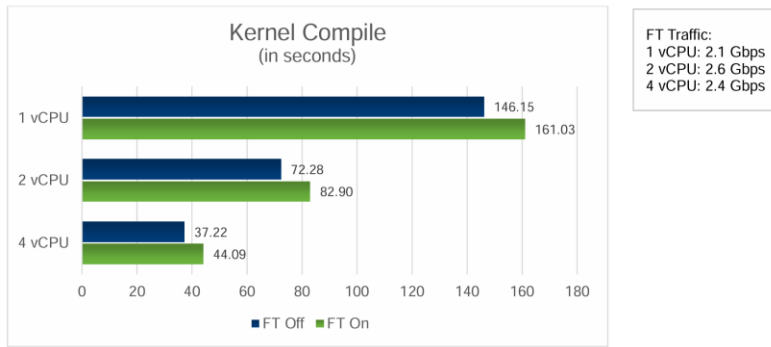


Figure 2. Kernel compilation performance (lower is better)

30

3. vSphere Fault Tolerance (FT)

VMware vSphere 6 Fault Tolerance Architecture and Performance
https://www.vmware.com/files/pdf/techpaper/VMware-vSphere6-FT-arch-perf.pdf

31

4. Virtual machine clustering

- 需求：
  - 某数据库业务使用 MS SQL Server，由 3 台服务器实现集群服务。
  - 数据库服务器使用 VM 实现（配置：16 vCPU、128G RAM、2 * 1GbE）
  - 需要实现数据库业务的高可用。

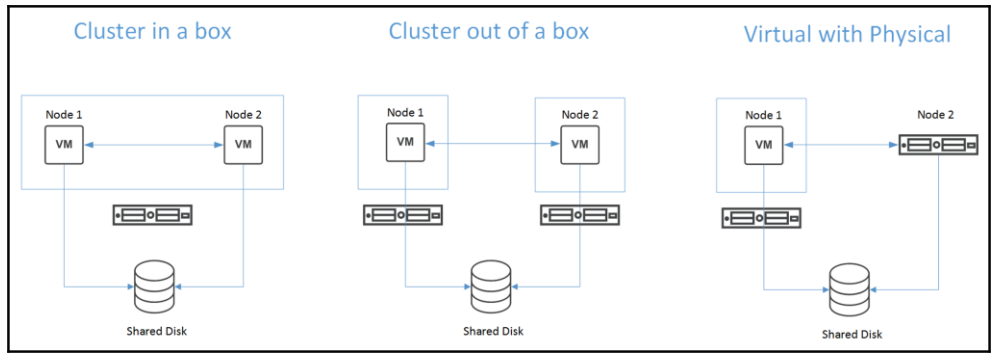
32

4. Virtual machine clustering

- Clustering VMs：对虚拟机进行聚类，虚拟机集群化
  - 将多个虚拟机组织在一起，形成一个具有特定特性和功能的集群。
  - 通过聚类来提高虚拟机的可靠性和可扩展性，实现业务高可用。
  - 具体的集群实现由操作系统和应用程序本身实现，和 vSphere 无关。
- vSphere 能够支持 Clustering VMs 的部署模式
  - vSphere 将多个系统和应用作为一个单一的逻辑单元来管理。
  - vSphere 支持三种模式的集群实现：
    - Cluster-in-a-Box ： VMs在同一 ESXi Host 实现集群。
    - Cluster-out-of-the-Box ： VMs 在多台 ESXi Host 实现集群。
    - VM and physical server clustering ： 推荐方案
      - 集群内的一个节点在物理服务器上运行。
      - 集群内的另一节点作为 VM 在 ESXi Host 上运行。

33

vSphere 支持 Clustering VMs 的部署模式



34

4. Virtual machine clustering

- vSphere 提供的 Clustering VMs 集群支持技术：
  - SCSI bus sharing for virtual disks on VMFS volume:
    - 多个 VMs 可以同时访问同一个虚拟磁盘。
  - SCSI bus sharing for RDM devices:
    - VM 不使用 VMFS 磁盘，直接映射 Raw 虚拟机格式。
  - Multi-writer flag on the virtual disk:
    - 多个 VMs 可以读写同一个虚拟磁盘。
  - In-guest iSCSI:
    - VM 不使用 VMFS 磁盘，使用映射的 iSCSI 存储.

35

4. Virtual machine clustering

- RDM：Raw Device Mapping，原始设备映射，直通式磁盘
  - RDM 允许一个虚拟机直接访问 SAN 中的一个存储 LUN（Logical Unit Number）。
  - RDM 磁盘实现虚拟机直接使用存储中的 LUN，而不经过虚拟化层。
- 使用RDM，VMkernel 不会对 LUN 进行格式化，而由虚拟机客户操作系统对 LUN 执行格式化。
  - 在存储效率和 IO 性能上，RDM 比创建的虚拟磁盘具有更好的表现。
  - RDM通常使用在高 IO 的 VM 上，例如：
    - Oracle、SQL Server 等 RDBMS
    - 用于视频点播或文件共享的大容量共享文件服务
    - Docker 或者 K8S 集群服务

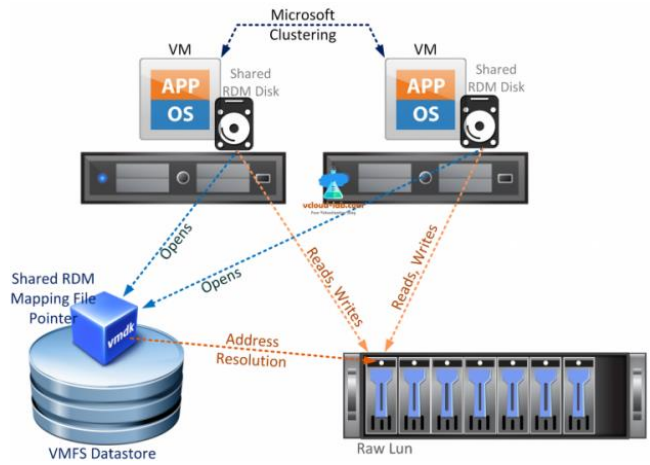
36

4. Virtual machine clustering

将 RDM 磁盘添加到虚拟机
https://docs.vmware.com/cn/VMware-vSphere/6.7/com.vmware.vsphere.vm_admin.doc/GUID-4236E44E-E11F-4EDD-8CC0-12BA664BB811.html

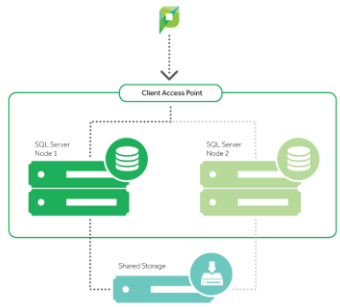
37

RDM 的工作流程



38

39



Configuring SQL Server Always On Failover Cluster Instances Using RDM On vSphere

什么是Always On可用性组件：
https://docs.microsoft.com/zh-cn/sql/database-engine/availability-groups/windows/overview-of-always-on-availability-groups-sql-server?redirectedfrom=MSDN&view=sql-server-ver15

部署指南：
https://tech2fun.net/configuring-sql-server-always-on-failover-cluster-instances-using-rdm-on-vsphere-7/

何为Always On：
https://blog.csdn.net/dba_huangzj/article/details/54015470

Getting Started with Always On Availability Groups：
https://docs.microsoft.com/en-us/sql/database-engine/availability-groups/windows/getting-started-with-always-on-availability-groups-sql-server?view=sql-server-ver15

39

40

5. VMware Solutions

5.1 VMware BC-related solutions

- 从传统业务向虚拟化迁移的最大驱动力是：业务连续性、业务高可用。
  - 虚拟化有助于服务器整合资源、利旧等，实现降低 IT 成本的目标。
  - 虚拟化更有助于提高可用性、业务弹性和故障恢复效率。
  - 影响业务连续性和高可用的因素有两个方面：
    - 非计划事件：服务器故障等
    - 计划事件：服务器维护等
- VMware 提供了整体的解决方案来提升可用性。
  - VMware BC-related solutions：
    - BC：关键业务与服务，business-critical applications and services
  - 方案具有简单易用、经济高效、部署灵活。

40

5. VMware Solutions

- VMware BC-related solutions:

本地可用：Local availability	vSphere HA，vSphere FT，vMotion…
数据保护：Data protection	Clone，Snap，Backup…
灾难恢复：Disaster recovery:	vSphere Replication，VMware SRM
灾难避免：Disaster avoidance	vSphere Metro Storage Cluster(vMSC)

41

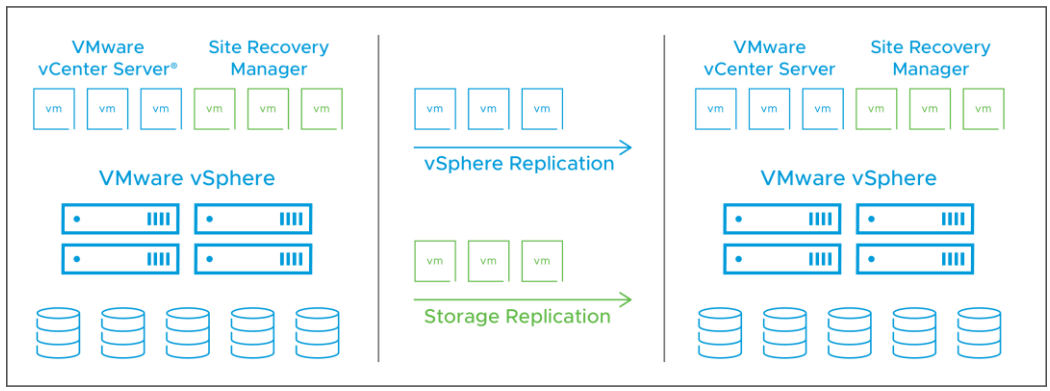
5. VMware Solutions

- VMware Site Recovery Manager:
  - 是一个业务连续性和灾难恢复解决方案，通过计划、测试和执行受保护 vCenter Server 站点与恢复 vCenter Server 站点之间的虚拟机恢复。
  - 通俗的说，就是 vSphere Replication 的自动化编排工具。
- VMware Live Recovery:
  - 将两个产品整合到一起，实现 One + One = More。
  - VMware Live Cyber Recovery
    - (formerly VMware Cloud Disaster Recovery + VMware Ransomware Recovery)
  - VMware Live Site Recovery
    - (formerly VMware Site Recovery Manager)
  - https://www.vmware.com/products/live-recovery.html

42

43

Site Recovery Manager automates the failover and migration of VMs to a secondary site.



43

44

5. VMware Solutions

5.3 vSphere Metro Storage Cluster (vMSC)

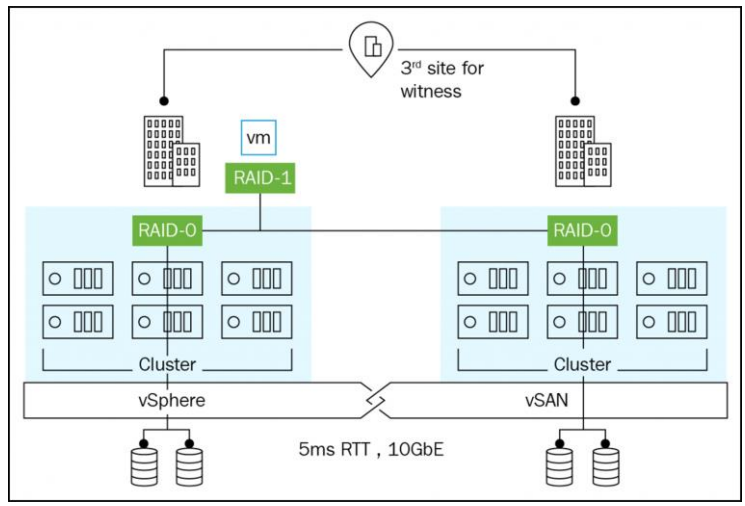
- VMware vSphere Metro Storage Cluster
  - VMware vSphere 城域存储集群，是一种高级的数据中心存储解决方案。
  - 能够在城域范围内的多个数据中心之间建立存储的协同工作机制，提供数据的同步复制和快速切换能力。
  - 实现跨不同地理位置的两个或多个数据中心之间的高可用性和灾难恢复能力。
  - 通过 vSAN 实现，利用了 vSAN 的分布式和拉伸集群功能，以构建一个单一的、逻辑上的存储资源池，这个资源池能够跨越多个数据中心。
- 通过 vSphere Metro Storage Cluster 可以：
  - 在不同的地理位置部署数据中心，当一个数据中心出现故障或面临灾难时，能够快速将业务切换到其他数据中心，保证业务的连续性和数据的安全性。
  - 与虚拟化技术紧密结合，为企业提供了强大的灾难恢复和业务连续性保障。

44

45

vSphere Metro Storage Cluster (vMSC)

双活



45

46

智能运维课程体系

46