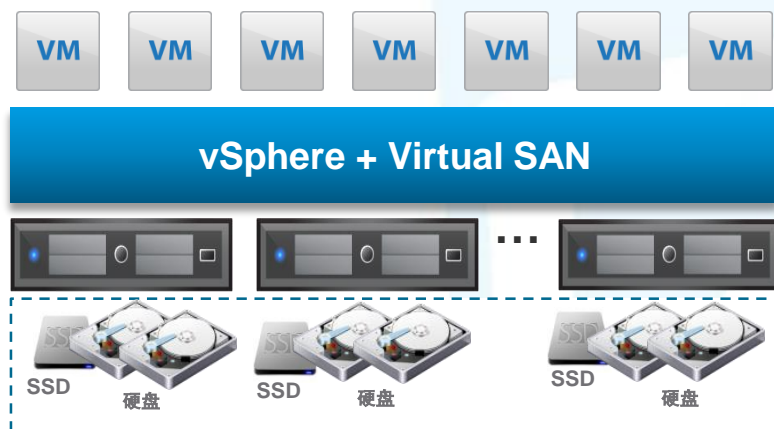


第十一节

VSAN

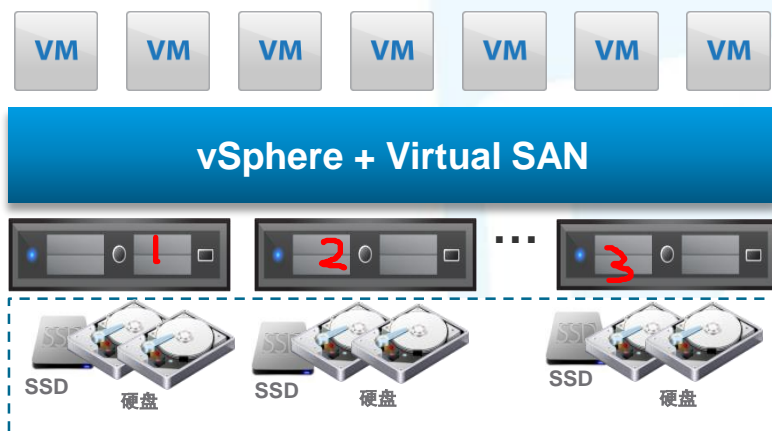
- 讲师:崔应龙
- 邮箱: cuiyl@5ibblue.com.cn

- 聚合了虚拟化管理程序的存储平台



- 软件定义的存储软件解决方案。
- 聚合集群中每台 ESXi 主机上本地连接的存储。
- 经过闪存优化的存储解决方案。
- 以虚拟机为中心的数据运营和策略驱动的管理原则。
- 基于分布式 RAID 体系结构的恢复能力强的设计。
 - 无单点故障
- 与 vSphere 全面集成。

- 聚合了虚拟化管理程序的极其简单的存储软件



Virtual SAN 共享数据
存储



- 混合式存储解决方案**
 - 磁盘（硬盘）
 - 基于闪存的磁盘（固态硬盘）
- 内置在虚拟化管理程序中的存储**横向扩展**体系结构
- 动态的**容量和性能可扩展性
- 基于对象的存储体系结构
- 可与 vSphere 以及下列企业级功能**互操作**:
 - vMotion、DRS、vSphere HA

基于 VMware 的任意
服务器兼容性指南



- VSAN集群至少 3 台 ESXi 6.0主机，最高64台主机

基于闪存的设备

在 Virtual SAN 中，**所有**读写操作都始终直接针对闪存层。

基于**闪存的设备**在 Virtual SAN 中具有两个作用

1. 非易失性**写缓冲区** (30%)

- 写入操作会在进入固态硬盘的准备阶段时确认
- 缩短写入延迟时间

2. **读缓存** (70%)

读一定大于写的操作

缓存命中说明读取的数据在缓存

- 缓存命中可缩短读取延迟时间
- 缓存未命中 – 需从硬盘检索数据

硬件选择是不同 Virtual SAN 配置之间的首要性能差异化因素。



确定闪存容量大小

- 一般情况下，建议将 Virtual SAN 闪存容量的大小定为未考虑容许的故障数时预计所用存储容量的 10%。

度量要求	值
预计虚拟机空间使用容量	20 GB(实际容量)
预计虚拟机数	1000
每个虚拟机预计占用的空间总量	$20\text{ GB} \times 1000 = 20,000\text{ GB} = 20\text{ TB}$
目标闪存容量百分比	10%
所需的闪存总量	$20\text{ TB} \times 0.10 = 2\text{ TB}$

- 总闪存容量百分比应该基于使用情形及其容量和性能要求来计算。
 - 10% 是一般建议，可能过多也可能不足。

例如：100台VM，每台VM设置100GB，预期平均为50GB

$10\% \times (100 \times 50\text{GB}) = 500\text{GB}$ SSD总容量，如果有5台主机，则每台SSD为100GB

VMware 固态硬盘性能级别

- A 级：每秒 2,500-5,000 次写入
- B 级：每秒 5,000-10,000 次写入
- C 级：每秒 10,000-20,000 次写入
- D 级：每秒 20,000-30,000 次写入
- E 级：每秒超过 30,000 次写入



工作负载定义

- 队列深度：16 个或更少 理解为缓存
- 传输长度：4 KB
- 操作：写入
- 模式：100% 随机
- 延迟：不到 5 毫秒

示例

- Intel 的 400 GB 910 PCIe 固态硬盘 每秒约 38000 次写入
- Toshiba 的 200 GB SAS 固态硬盘 MK2001GRZB 每秒约 16000 次写入

闪存攻克关键

耐久性

- 10 次驱动器写入/天 (DWPD)，以及
- 每个 NAND 模块的传输长度为 8 KB 时最多 3.5 PB 随机写入耐久性，而每个 NAND 模块的传输长度为 4 KB 时最多 2.5 PB

- 支持的 SAS/NL-SAS/SATA 硬盘
 - 7200 RPM 用于提供容量
 - 10000 RPM 用于提供性能
 - 15000 RPM 用于提供更高性能
- 在驱动器转速相同和价位相似的情况下，NL SAS 将提供更高的硬盘控制器队列深度
 - 如果在 SATA 与 NL SAS 之间选择，建议选择 NL SAS
- 选择不同的固态硬盘以及不同的固态硬盘与硬盘比率，会使集群性能有所区别。按照经验法则，应按 10% 的比率配置



- SAS/SATA 存储控制器 raid卡

- 支持直通或“RAID0”模式



- 使用 RAID0 模式时，性能取决于控制器
 - 请咨询您的供应商以了解 RAID 控制器的固态硬盘性能
- 存储控制器队列深度很重要
 - 加大存储控制器队列深度将可以提高性能

这的队列深度可理解为缓存
- 核实每种控制器支持的驱动器数量

存储控制器 – RAID0 模式

- 将所有磁盘配置为 RAID0 模式
 - 基于闪存的设备（固态硬盘）
 - 磁盘（硬盘）
- 禁用存储控制器缓存
 - 可实现更高性能，因为缓存由 Virtual SAN 控制
- 磁盘设备缓存支持
 - 基于闪存的设备利用直写缓存
 - 磁盘利用写回缓存
- ESXi 可能无法区分基于闪存的设备与磁性设备
 - 使用 ESXCLI 手动将设备标记为固态硬盘





- 支持 1 GB/10 GB
 - 带确保服务质量的 NIOC 的 10 GB 共享网络将支持大多数环境
 - 如果是 1 GB 网络，则建议对 Virtual SAN 使用专用链路
- 巨型帧将提供标称的性能提升
 - 针对全新部署启用
- Virtual SAN 同时支持 VSS 和 VDS
 - NIOC 需要使用 VDS
 - Nexus 1000v – 应该能用，但尚未进行全面测试
- 网络带宽性能对主机撤出、重建时间的影响高于对工作负载性能的影响

MTU最大传输单元

标准交换机

分布式交换机

防火墙

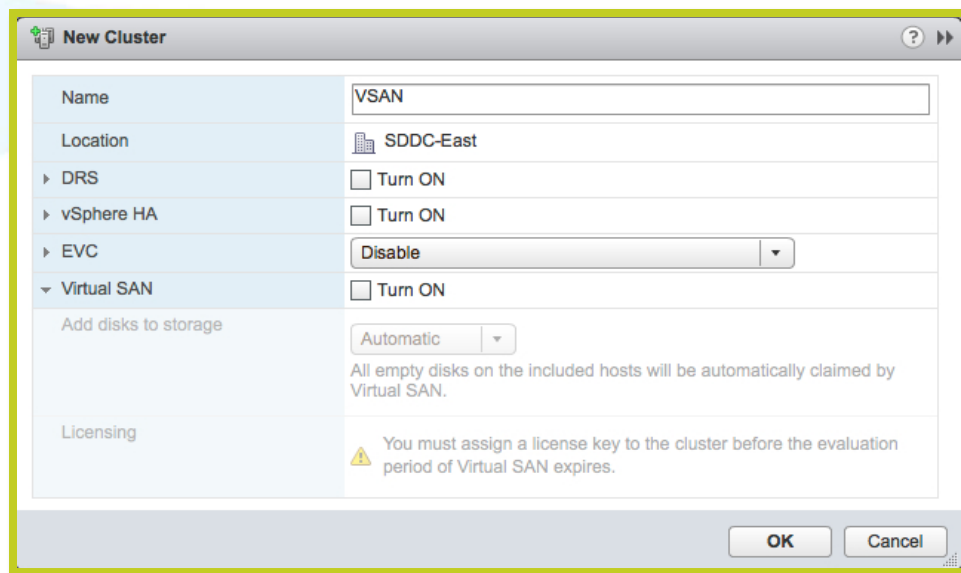


- Virtual SAN 供应商提供程序 (VSANVP)
 - 入站和出站 - TCP 8080
- 集群监控、成员资格和监控服务 (CMMDS)
 - 入站和出站 UDP 12345 - 23451
- 可靠数据报传输 (RDT)
 - 入站和出站 TCP 2233

配置--安全配置文件--中设置防火墙,入站与出站的配置,关于安全的

Virtual SAN 是集群级别的功能，类似于：

- vSphere DRS
- vSphere HA
- Virtual SAN



通过 vSphere Web Client 从 vCenter 中进行部署、配置和管理

(绝无仅有!)。

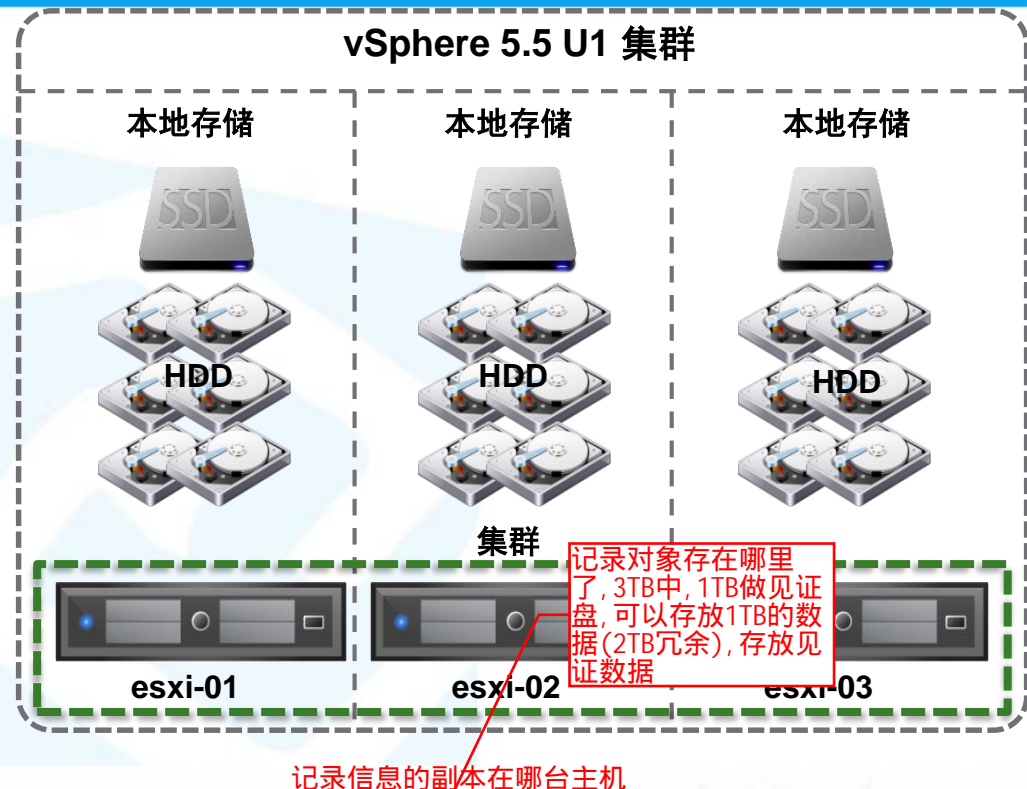
- 极其简单

- 为 Virtual SAN 配置 VMkernel 接口
- 通过单击“Turn On”（打开）启用 Virtual SAN

直接集成到了
vcenter中, 要想搭建
vcenter只能在web端

Virtual SAN 实施要求

- Virtual SAN 需要：
 - 采用集群配置，至少 3 台主机
 - 3 台主机都**必须**提供存储
 - vSphere 5.5 U1 或更高版本
 - 本地连接的磁盘
 - 磁盘（硬盘）
 - 基于闪存的设备（固态硬盘）
 - 网络连接
 - 1 GB 以太网
 - 10 GB 以太网（首选）



允许的故障数	镜像 / 副本	见证对象	最少 ESXi 主机数
0	1	0	1
1	2	1	3
2	3	2	5
3	4	3	7

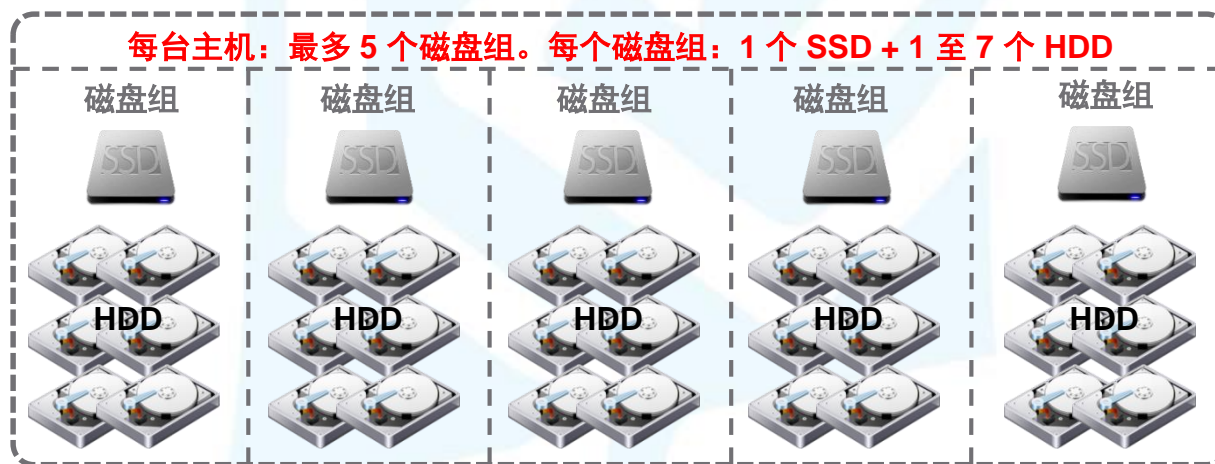
如果在虚拟机部署的时候没有选择任何策略，默认策略会将允许的故障数设成 1。在创建一个新策略的时候，允许的故障数的默认值也是 1。这意味着即使没有在策略中明确说明这个功能，它也已经暗含在内了。

全新的 Virtual SAN 构造、项目和术语：

- 磁盘组
- VSAN 数据存储
- 对象
- 组件
- Virtual SAN 网络

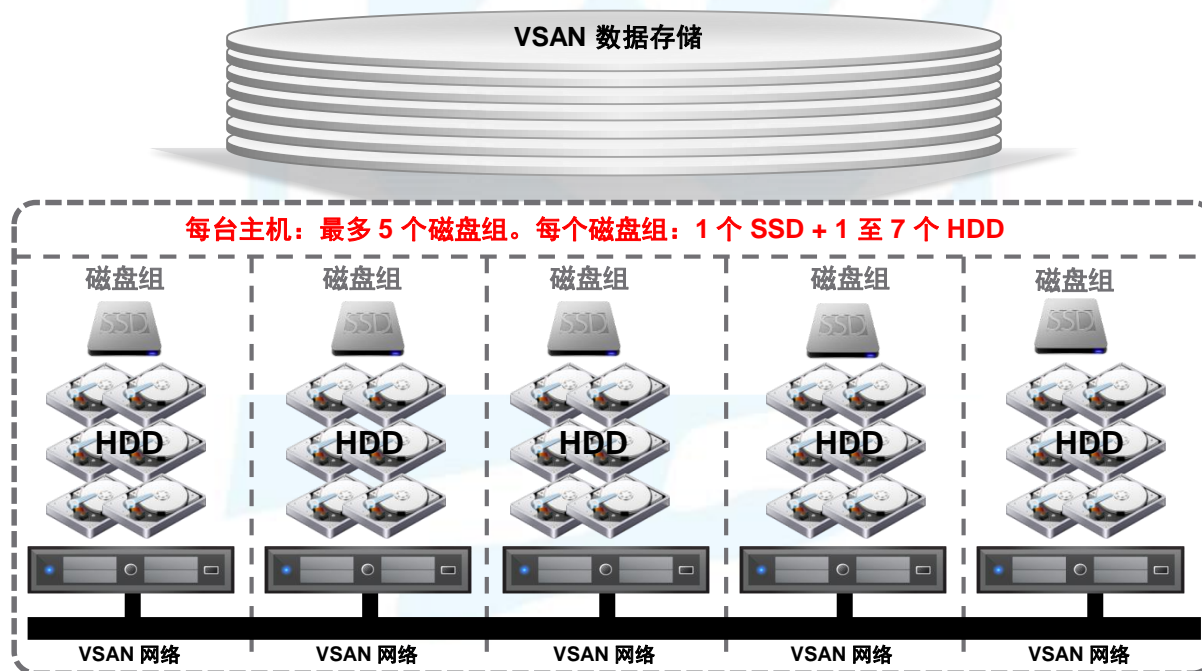
Virtual SAN 磁盘组

- Virtual SAN 使用**磁盘组**这一概念将闪存设备和磁盘池化为一个管理构造。
- 磁盘组至少包含 **1 个闪存设备**和 **1 个磁盘**。
 - 闪存设备用于提供性能（读缓存 + 写缓冲区）。
 - 磁盘用于提供存储容量。
 - 不能**在没有闪存设备的情况下创建磁盘组。



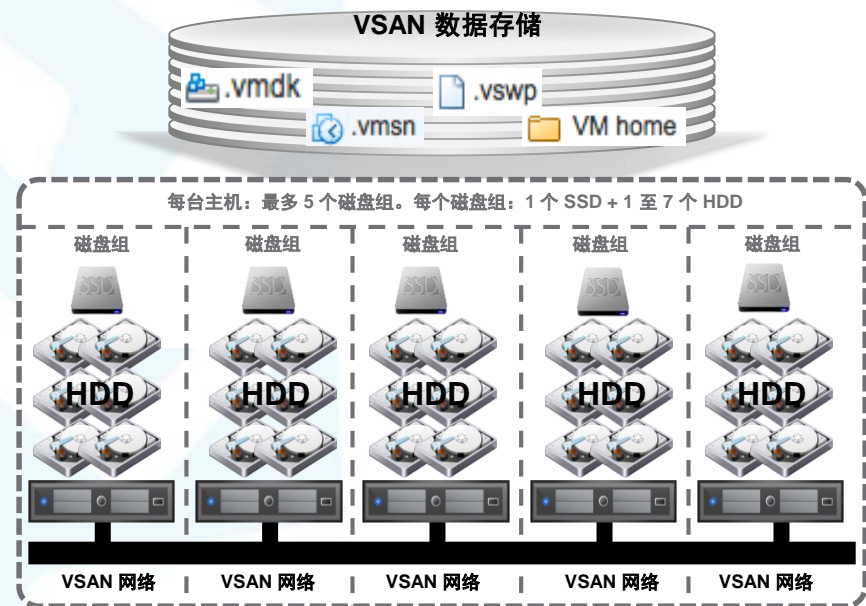
Virtual SAN 数据存储

- Virtual SAN 是一种以文件系统的形式呈现给 vSphere 的对象存储解决方案。
- 该对象存储装载着集群中所有主机的 VMFS 卷，并将它们呈现为一个共享数据存储。
 - 仅限该集群的成员才能访问 Virtual SAN 数据存储。
 - 并非所有主机都需要提供存储，但是建议提供存储。
 - 每个VSAN群集的数据存储数量是1个



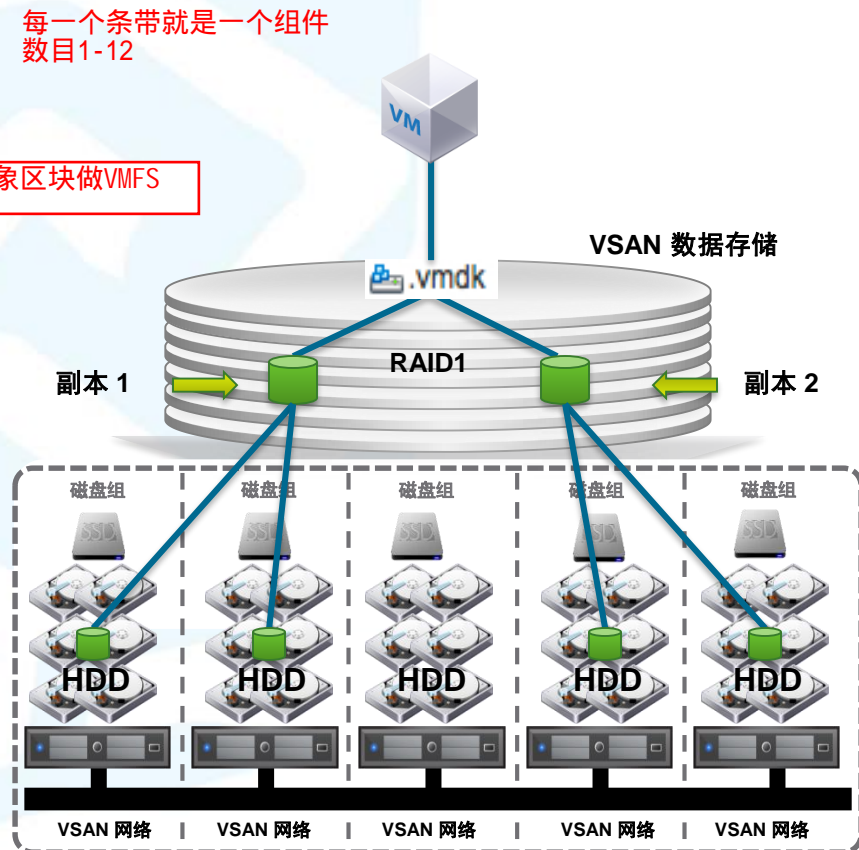
Virtual SAN 对象

- Virtual SAN 通过名为**对象**的灵活数据容器的形式管理数据。虚拟机文件即称为对象。
- 虚拟机文件称为对象。
 - 存在四种不同类型的虚拟机对象：
 - 虚拟机主目录
 - 虚拟机交换文件
 - VMDK
 - 快照
- 虚拟机对象基于虚拟机存储配置文件中定义的**性能**和**可用性**要求划分为多个**组件**。



Virtual SAN 组件

- Virtual SAN 组件是**对象区块**，这些对象区块跨集群中的多台主机分布，以便容许同时发生多个故障并满足性能要求。
- Virtual SAN 利用**分布式 RAID** 体系结构将数据分发到整个集群中。
- 组件的分布主要采用两种技术：
 - 条带化 (RAID0)
 - 镜像 (RAID1)
- 创建多少组件副本将基于对象策略定义决定。

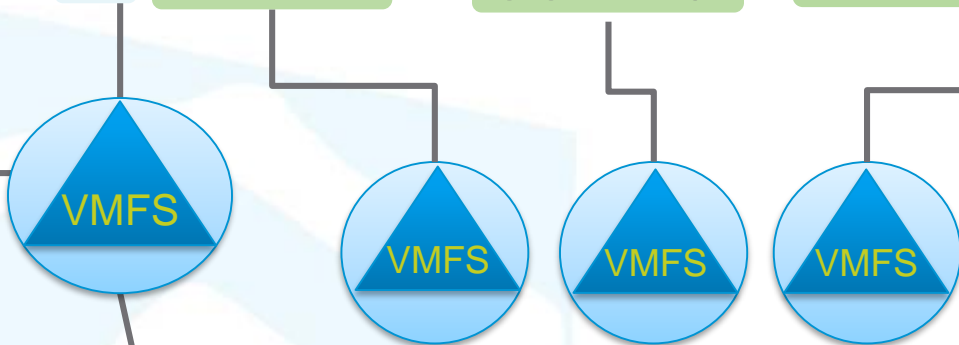


对象和组件布局

/vmfs/volumes/vsanDatastore/rolo/rolo.vmdk rolo1.vmdk rolo2.vmdk

rolo.vmx、.log 等

虚拟机主目录对象格式化为 VMFS，
以便在此对象上存储虚拟机的配置
文件。



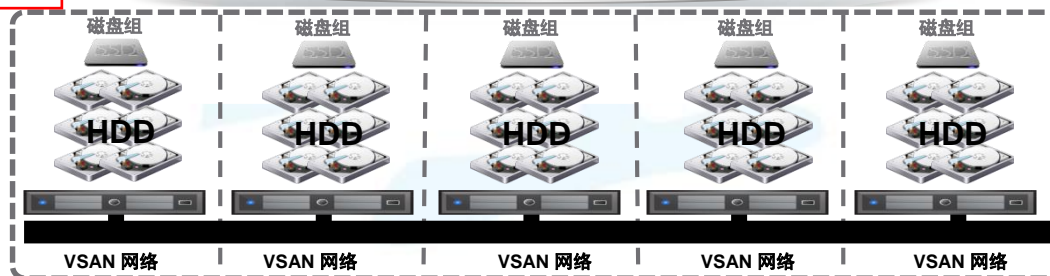
Virtual SAN 存储对象

镜像

可用性定义为副本数量

raid 0条带

性能可能会包括条带宽度

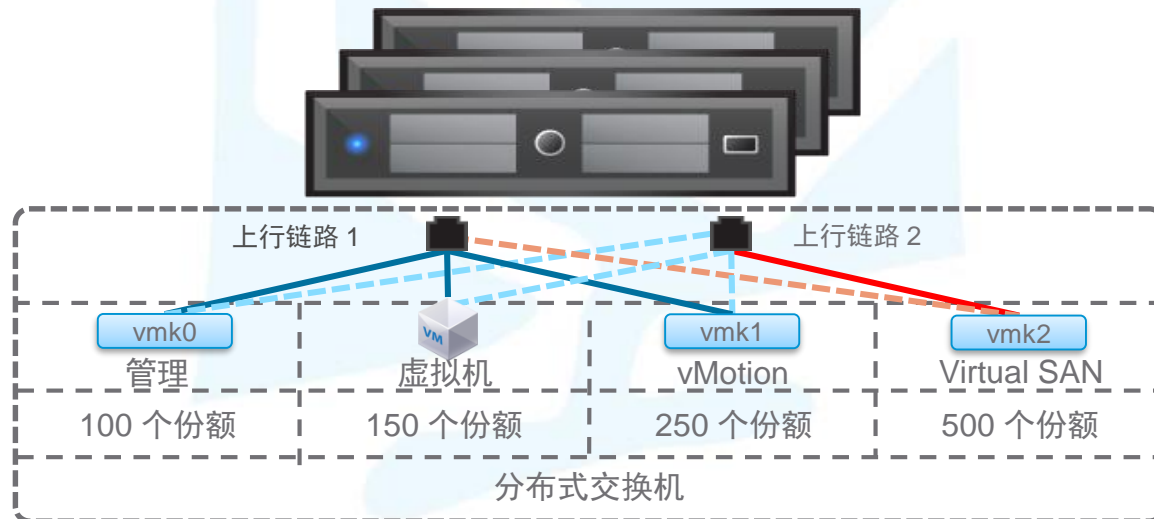


低级存储对象
将驻留在不同的
主机上

Virtual SAN 网络

- 新的 Virtual SAN 流量 VMkernel 接口。
 - 专供 Virtual SAN 进行**集群内**通信和数据复制使用。
- **同时**支持标准和分布式虚拟交换机
 - 在共享场景中利用确保服务质量的 NIOC
- 网卡绑定 – 用于**提高可用性**而不是用于带宽聚合。
- **第 2 层**多播**必须**在物理交换机上启用。
 - 比第 3 层多播更易于管理和实施

冗余容灾



• 网卡绑定和负载均衡算法：

— 基于端口 ID

- 路由**主动/被动**，采用显式故障切换

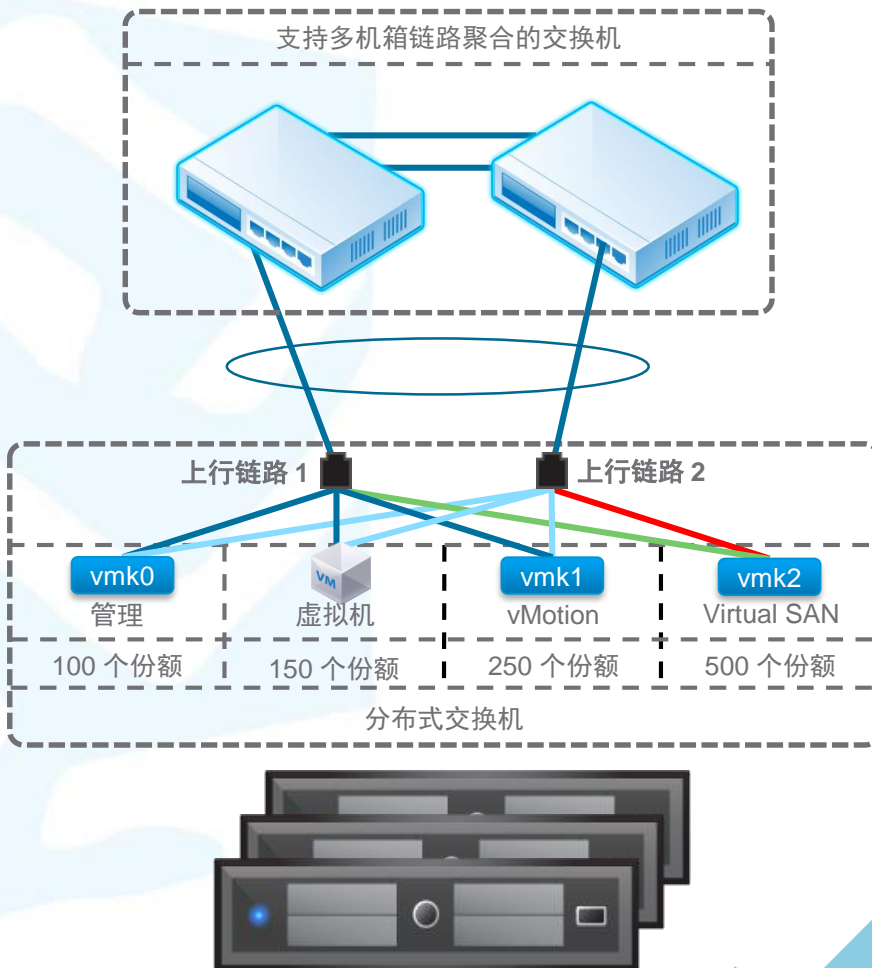
— 基于 IP 哈希

- 路由**主动/主动**，采用 LACP 端口通道

— 基于物理网卡负载

- 路由**主动/主动**，采用 LACP 端口通道

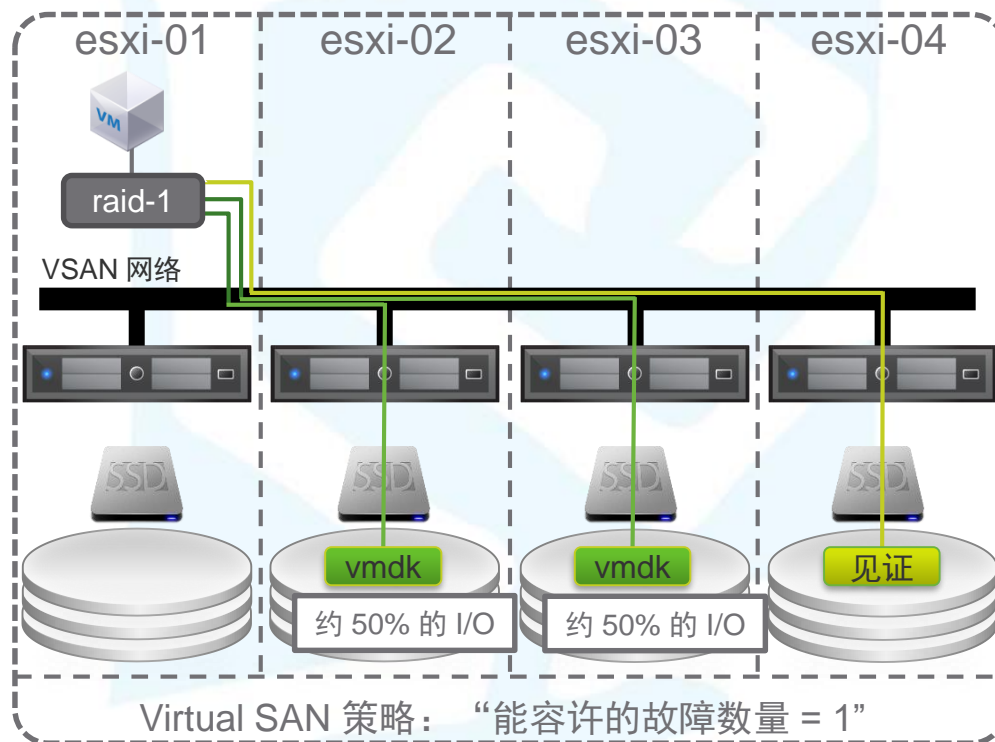
网络--负载均
衡---必须在外边物
理机上做以太网络端
口聚合



容许的故障数量

- 容许的故障数量

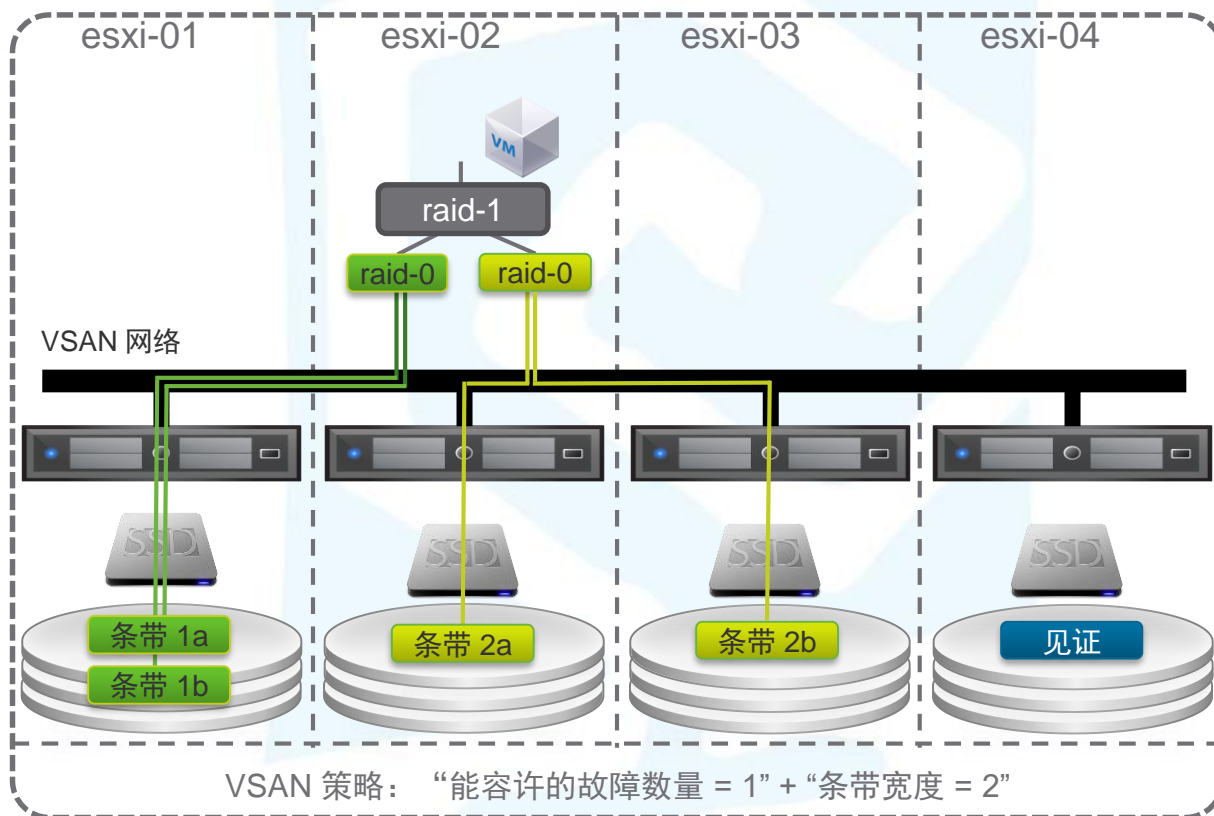
- 定义存储对象能容许的主机、磁盘或网络故障的数量。若要容许“ n ”个故障，则要创建“ $n+1$ ”个对象副本，并且需要“ $2n+1$ ”台主机提供存储。

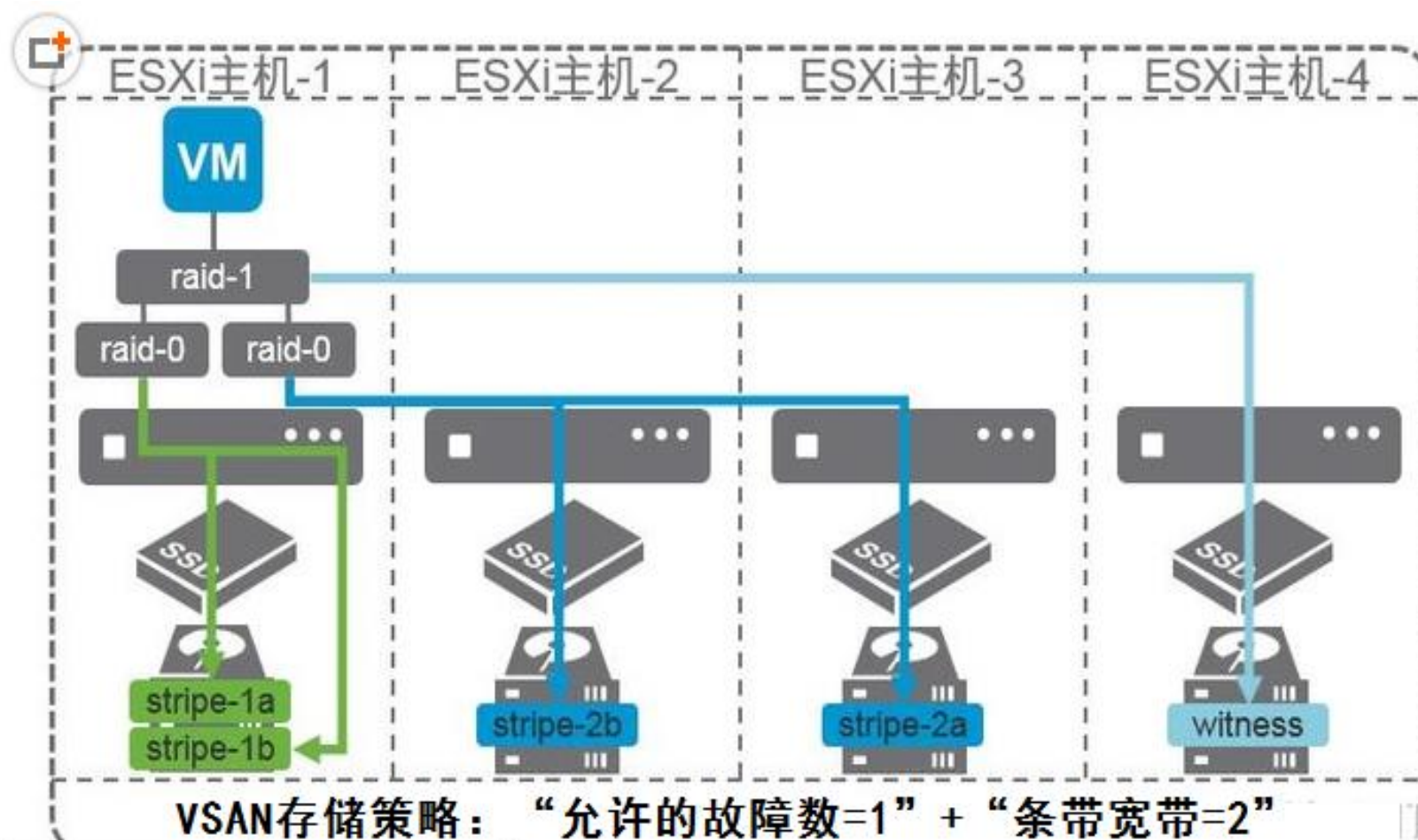


每个对象的磁盘条带数

- 每个对象的磁盘条带数

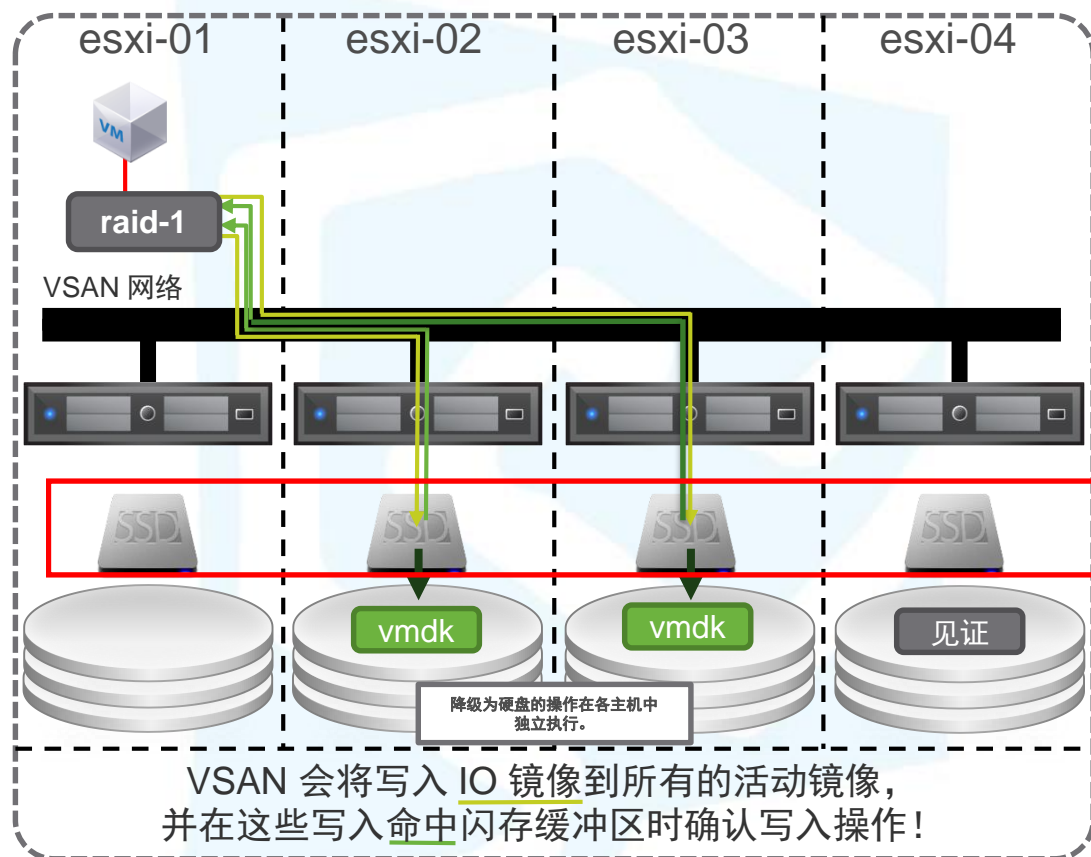
- 存储对象的每个副本所跨的硬盘数。值越高，性能就越好。



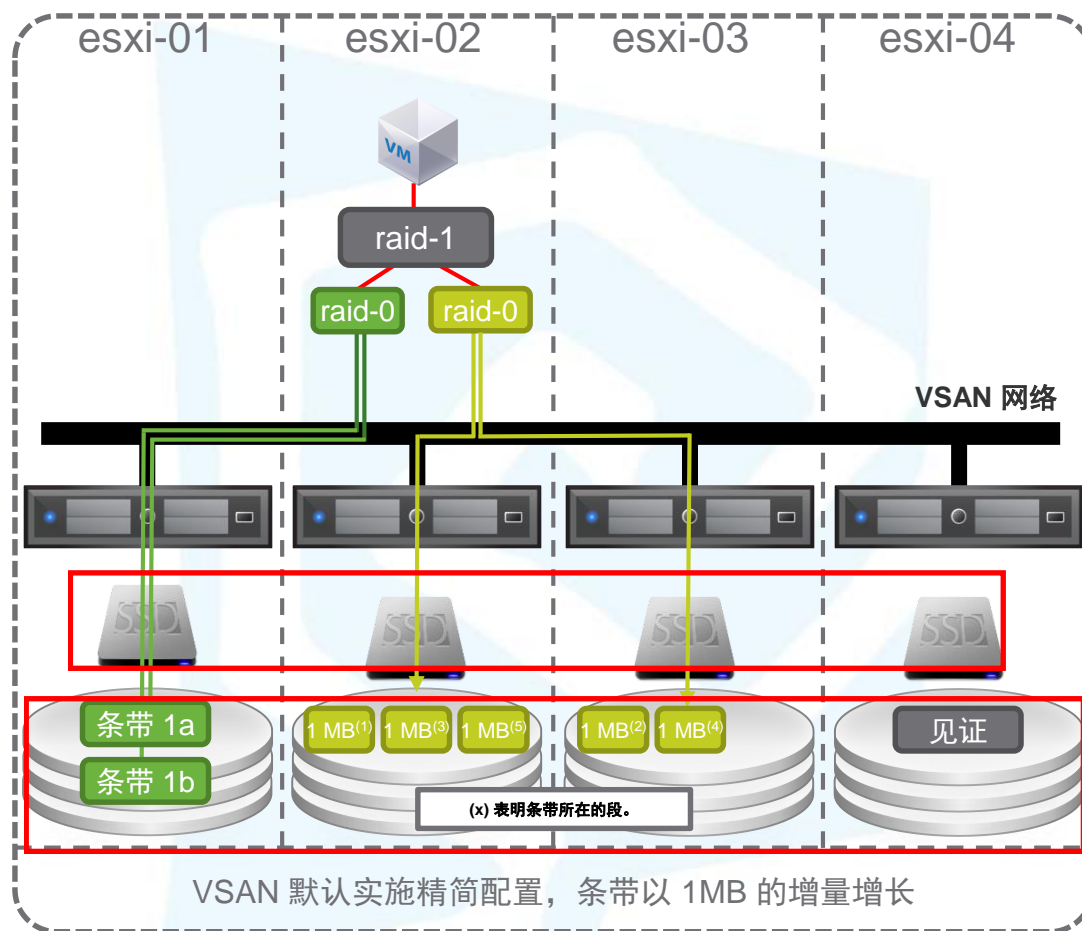


- 强制调配
 - 如果选择“**Yes**”（是），则即使当前可用的资源不符合存储策略中指定的策略，仍将调配对象。
- 闪存读缓存预留 (%)
 - 预留闪存容量，作为存储对象的读缓存。以对象逻辑大小的百分比形式指定。
- 对象空间预留 (%)
 - 调配虚拟机时要预留（实施厚配置）的存储对象的逻辑大小的百分比。将对其余存储对象实施精简配置。

Virtual SAN I/O 流 – 写入确认







Virtual SAN I/O 流 – 1 MB 增量的条带化



针对存储功能的建议做法

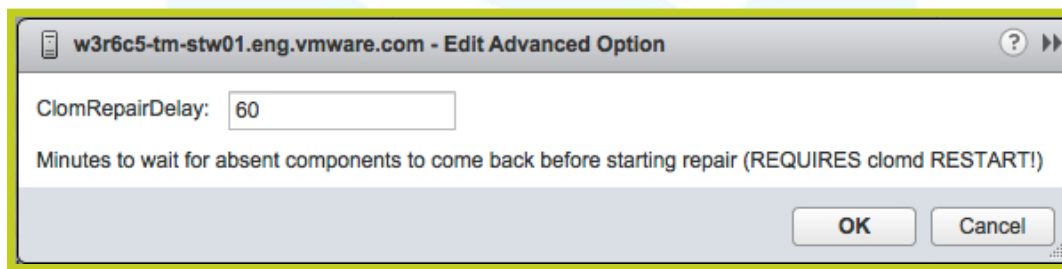
存储功能	使用情形	值
容许的故障数量 (RAID 1 – 镜像)	冗余	默认 1 最大 3
每个对象的磁盘条带数 (RAID 0 – 条带)	性能	默认 1 最大 12
对象空间预留	厚配置	默认 0 最大 100%
闪存读缓存预留	性能	默认 0 最大 100%
强制调配	覆盖策略	禁用

虚拟机存储策略建议

- 每个对象的磁盘条带数 
 - 应该保留为 1，除非闪存层未满足虚拟机的 IOPS 要求。
- 闪存读缓存预留 
 - 应该保留为 0，除非虚拟机要满足特定的性能要求。
- 比例容量 
 - 应该保留为 0，除非需要虚拟机厚配置。
- 强制调配 
 - 应该保留禁用，除非需要调配虚拟机（即使不合规）。

了解故障事件

- Virtual SAN 可识别**两种**不同类型的硬件设备事件以便定义故障场景的类型：
 - 缺失
 - 降级
- 缺失事件**会触发 60 分钟的恢复操作。
 - Virtual SAN 会在开始恢复对象和组件之前等待 60 分钟
 - 60 分钟是所有缺失事件的默认设置
 - 此值可通过主机高级设置来配置



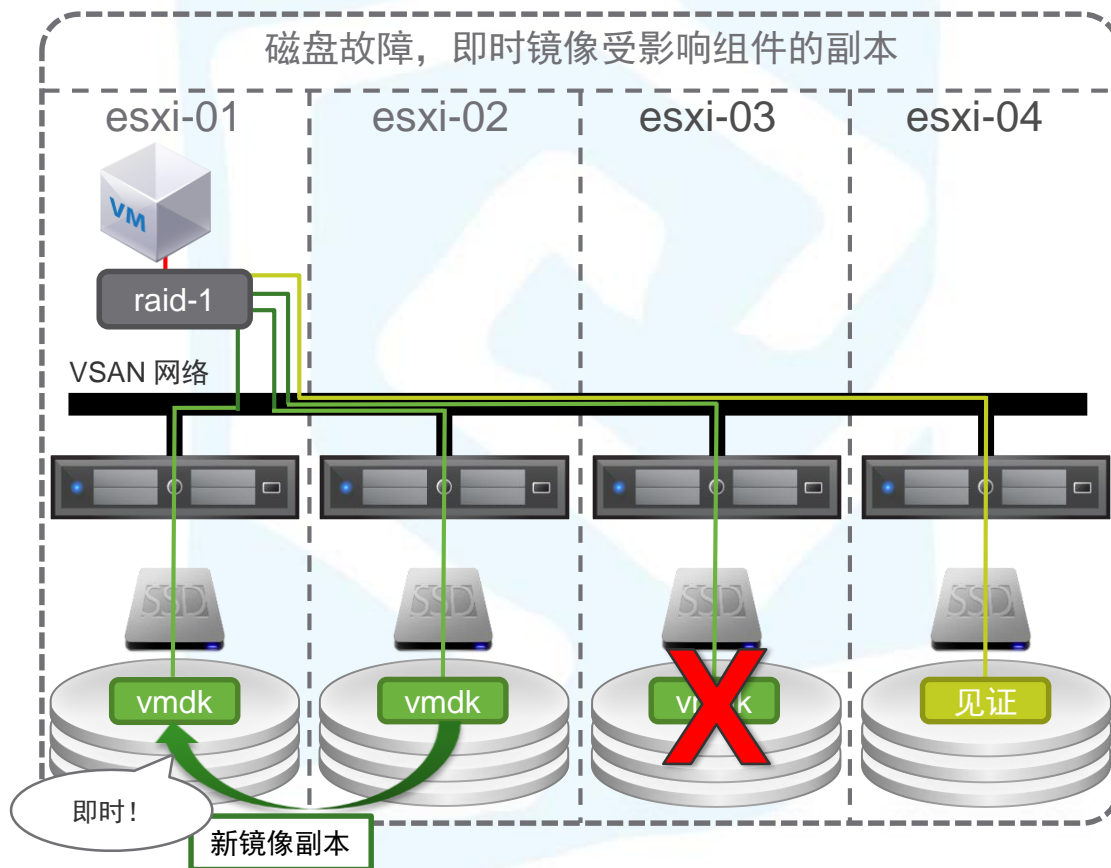
了解故障事件

- **降级事件**会立即触发恢复操作。
 - 立即触发对象和组件的恢复操作
 - 不可配置
- 如检测到的以下任何 I/O 错误始终会被视为**降级**:
 - 磁盘故障
 - 基于闪存的设备故障
 - 存储控制器故障
- 如检测到的以下任何 I/O 错误始终会被视为**缺失**:
 - 网络故障
 - 网卡 (NIC)
 - 主机故障

- 通过制定策略，Virtual SAN 上的虚拟机可以容许多种故障
 - 磁盘故障 – 降级事件
 - 固态硬盘故障 – 降级事件
 - 控制器故障 – 降级事件
 - 网络故障 – 缺失事件
 - 服务器故障 – 缺失事件
- 虚拟机可继续运行
- 并行重建可最大限度减少性能影响
 - 固态硬盘故障 – 立即
 - 硬盘故障 – 立即
 - 控制器故障 – 立即
 - 网络故障 – 60 分钟
 - 主机故障 – 60 分钟

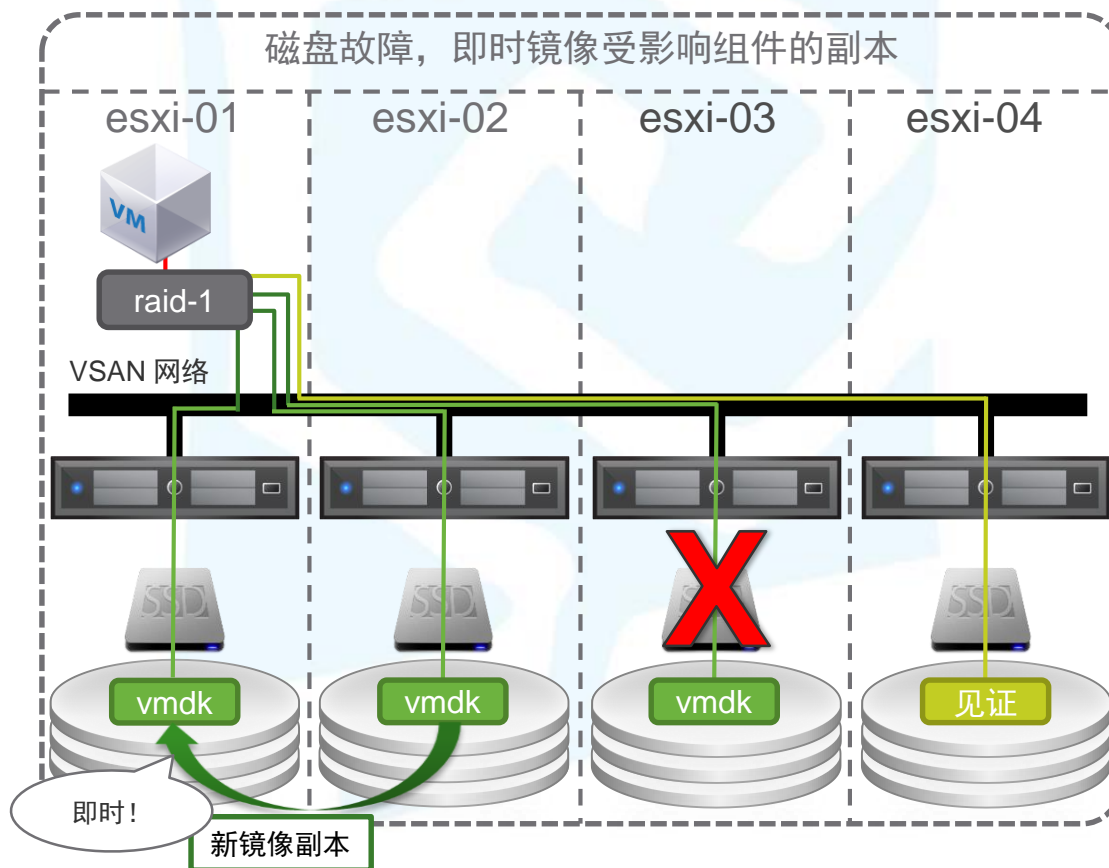
磁盘故障 – 即时镜像副本

- **降级** – 故障磁盘上所有受影响的组件都将立即在其他磁盘、磁盘组或主机上创建。



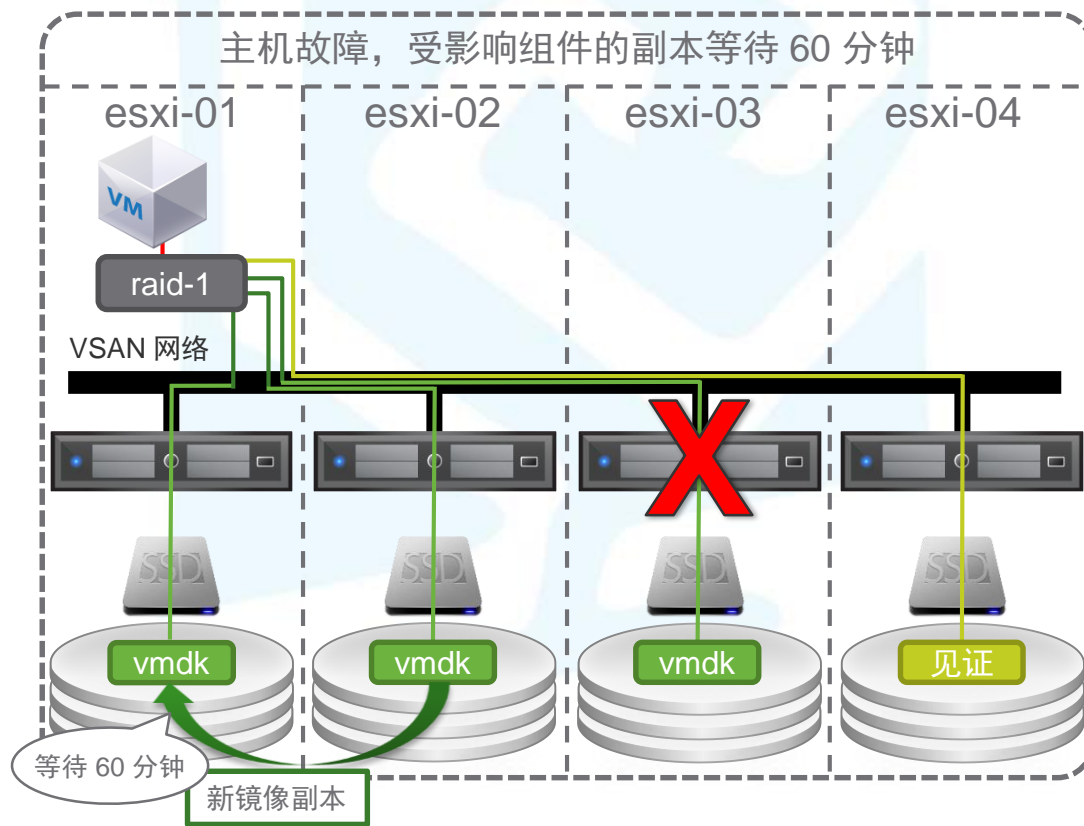
基于闪存的设备故障 – 即时镜像副本

- **降级** – 故障磁盘上所有受影响的组件都将立即在其他磁盘、磁盘组或主机上创建。
- 对集群总体存储容量的影响较大



主机故障 – 60 分钟延迟

- **缺失** – 在其他磁盘、磁盘组或主机上启动对象和组件的副本之前，将按默认设置等待 60 分钟。
- 对集群总体计算和存储容量的影响较大。



网络故障 – 60 分钟延迟

- **缺失** – 在其他磁盘、磁盘组或主机上启动对象和组件的副本之前，将按默认设置等待 60 分钟。
- 网卡故障、物理网络故障可能导致网络分区。
 - 可能会影响集群中的多台主机。

