日志文件
更多阅读

> 此文章假设读者已有基础VSAN知识或实施经验。
> 写了一半，感觉完全是在写VSAN考试大纲...
> 花了两个月时间学习并总结这个文档，期间VSAN项目中也遇到不少问题，学习完这个文档才发现VSAN并没有想象中的那么简单，所以非常建议做VSAN设计的人员能学习下此文章(或原文档)，避免规划出现问题。

概要

此文章遵照原参考文档的目录，着重介绍了存储、存储策略、主机和集群的设计。

- 其中存储部分包括磁盘组设计、缓存盘(SSD，nvme，PCIE闪存盘)和HDD（SAS，SATA，NL-SAS）选择、VSAN文件系统的开销、RAID卡的设计等。
- 存储策略着重介绍了当前VSAN 6.2的所有策略功能及对物理存储使用情况的介绍，以及默认VSAN中虚拟机目录、交换文件、快照等的存储占用情况。
- 主机着重介绍VSAN下启动设备的选择、刀片服务器支持、外置磁盘柜支持、电源管理等。
- 集群着重介绍2/3 VSAN节点的使用、HA、故障域及去重、压缩
- 最后为两个较为详细的配置案例，其中纠正了原文档中的一些小错误

简介

Virtual SAN 是一个集成在Hypervisor中，与VMware vsphere高度集成的软件定义存储产品。VSAN可以将**一个**vSphere cluster集群中主机上的**直连硬盘**整合起来，创建出**一个**分布式共享存储。它使用策略(文中会有策略的详细描述)驱动，简化了存储的置备和管理。

VSAN 当前有两种配置选项，混合配置(hybird)及全闪配置(all-flash)；混合配置下一个磁盘组使用一个SSD做读写缓存(容量7/3划分)，机械硬盘存储数据。全闪模式下使用寿命高的SSD作为写缓存盘(100%空间用于写缓存，最大只能使用600G空间)，使用读性能较好、寿命不高但价格低廉的SSD作为数据存储盘。

> 解读：
> 每个VSAN存储对应一个集群，一一对应，一个集群中所有主机都会开启VSAN功能（需要手动配置VSAN网络），一个集群只能有一个VSAN，所有硬盘都在这一个VSAN中，VSAN中的存储不能直接供其他集群直接使用，需要通过创建NAS存储的方式对外提供存储服务。
> VSAN 基于 storage policies，通过策略可以定义每个虚机副本数(以此决定主机/硬盘最大故障数)，还有资源预留、条带等很多参数设置。

开始之前

- 请通过VMware兼容性指南认真检查所使用的服务器、RAID卡、PCIE flash、SSD、机械硬盘是否

在兼容列表内。

- 请确保使用的软件、驱动及固件在兼容列表内，兼容列表内还会提供最佳的驱动、固件版本号。
- 在执行全新部署前，确保vSphere安装了所有补丁程序。可以考虑对已有部署进行升级。
- 建议群集中所有主机使用相同的配置。
- VSAN 可以通过添加硬盘实现纵向扩容，添加主机实现横向扩容。在进行扩容时要考虑到缓存/容量盘的容量比。**更换SSD硬盘会导致一整台主机上的所有数据重建**，所以前期最好规划好SSD容量。而HDD添加/更换影响的只是单个盘上的数据。
- VSAN最少需要三台ESXi主机或者两个ESXi主机加一个VSAN见证。当有主机故障时，VSAN会自动尝试修复虚拟机的文件，一个对象至少两份数据一个见证共三个组件，因此三台主机的VSAN环境如果有一台ESXi故障，那么其中损坏的组件是无法被恢复的。主机处于维护模式时，为保证虚拟机数据正常，需要将维护模式主机上的数据迁移到其他可用ESXI上，三台主机VSAN环境也不能实现这点。
- VSAN 通过策略实现虚拟机数据保护，其中一个策略为FTT(允许的故障数)，此策略控制虚拟机数据存放的副本数。在进行VSAN设计时，需要根据虚拟机冗余程度确定整体容量。

VSAN 限制

- VSAN最少需要三台ESXi主机，最大支持64台主机
- VSAN 6.0 每台主机上可以存放200台虚拟机，整个集群最大支持6400个虚拟机。
- VSAN 6.0 每个集群最大可以通过HA保护6400个虚拟机
- 每台VSAN主机最大可有5个磁盘组，每个磁盘组最大可以有7个容量磁盘，一个缓存磁盘
- 虚拟机在VSAN上以对象的形式存在，例如一个VMDK文件就是一个对象，一个快照文件就是一个对象，VM目录是一个对象。这每一个对象都由多个组件(component)组成，组件的数量由存储策略决定，例如FTT=1时，会有两份component。VSAN 5.5 最大支持3000各组件。VSAN 6.0每台主机最大支持9000个组件。
- 每个对象的最大条带数是12，最小是1，但是有些情况下存储策略中没有设置条带时，VSAN也会自动设置一个对象的条带。原因有多种，例如VSAN规定一个组件最大为255G，如果一个对象(vmdk)大于255G时，VSAN自动将这个对象拆成多个组件。因此当有一个2T的vmdk文件时，它很可能会由8个或者更多个组件组成RAID-0(FTT=1时，一个vmdk会包含一个witness，两个RAID-0)。 在设计时，需要根据设置的条带数确定硬盘数够不够。
- FTT 最大为3，默认存储策略FTT=1
- FlashReadCacheReservation 最大为100%，表示一定会给vm预留匹配大小的缓存，此策略只适用于混合模式VSAN
- ObjectSpaceReservation 最大100%，表示虚拟机会被置备为厚模式。
- VSAN 6.0最大VMDK支持64T，但是每个对象依然有组件大小255G的限制

- 混合模式下网络支持1G和10G，当使用1G时需要独占网卡，10G时可以让VSAN流量和其他流量混跑。全闪存只支持10G及以上链路，也可以让VSAN流量和其他流量混跑，但是建议使用Qos限制每种流量的带宽占用。
- VSAN不支持多链路负载均衡，多链路只能实现高可用（即使使用了LACP）。
- 使用Jumbo Frame可以减少CPU负担改善带宽，但是效果很小，因为vSphere有TSO和LRO （更多信息请看KB2055140）提供类似的功能。另外不建议在Jumbo Frame支持不好的交换机上开启（例如开启后交换机CPU利用过高）
- 全闪模式下建议最好开启Jumbo Frame
- 交换机必须支持组播功能，建议采用高端的企业级交换机
- 可以使用NIOC(Network I/O control)进行Qos，对VSAN流量进行带宽保证或者限制。此功能仅支持VDS。

一、网络设计

另起文章：VSAN 6.2 网络设计

一句话总结：**保持一个VSAN集群在一个二层网络内，之间使用全线速交换机(交换机堆叠)连接。**

二、存储设计

2.1 磁盘组

> 关键词：磁盘组数量及容量，同时考虑未来纵向扩展(加磁盘组)

磁盘组可以看做VSAN的容器，一个磁盘组至少包含一个SSD和一个HDD。如果希望提高缓存和容量比，可以在一个主机上**创建多个磁盘组**，多个磁盘组可以提高IOPS，减少故障域(即，在相同缓存容量比下，单个磁盘组SSD一定是大容量，IOPS有上限，且当此SSD损坏时，相当于整个主机故障。多个磁盘组用多个SSD，IOPS有可能会更高，且随意一个SSD故障，只影响所在磁盘组的数据)。

2.2 缓存盘设计

> 关键词：容量(考虑未来使用及未来纵向扩容)，寿命

缓存大小应该根据工作中热点数据的量决定，但是这个值不好统计，可能随着时间变化而不同，VMware建议缓存/容量比至少为1:10（**混合模式及全闪模式均适用，且计算方法一样**）。

缓存的计算应该不考虑FTT，按照实际使用容量计算，例如：1000台虚拟机，每台瘦置备100G空间，但是平均使用空间为20GB，那么总使用容量为1000*20G=20T，按照1:10的比例，总闪存盘大小为2TB，如果是4个节点VSAN，则每个主机可能要安装600G的SSD。但是，考虑到未来增长，可能VM实际占用空间会增加到30G、40G，这样缓存盘就需要更大的了，前期一定要做好一定量的预留。

全闪存VSAN下，每个磁盘组的缓存盘最大只能使用600G(All-flash自身限制，因此容量过大就是浪费)，全被作为写缓存。写缓存可以减少容量层SSD的写操作，延长这些磁盘的寿命。在全闪模式下，一定要注意flash(缓存)层磁盘的寿命！

2.2.1 读缓存

读缓存仅在混合模式有效。VSAN将vm最近读取的数据块保存在闪存中，减少读取延迟，提高读取效率。

当有VM有多个副本时，VSAN均衡地从每个副本读取数据块放在SSD中。

如果需要读取的数据块不在一个SSD中，VSAN directory service会在其他主机的SSD中搜索此块，如果最终未找到，则出现一个read cache miss，数据直接从磁盘去读取。

2.2.2 写缓存

写缓存在混合模式和全闪存模式中都存在，可以增加性能，提高全闪存时容量层SSD的寿命。

当OS中有应用发起写入请求时，此请求会被复制到存有此对象的主机的SSD上，因为SSD是非易失性存储，所以在一个主机故障时不会引起数据丢失。

混合配置下，SSD上的数据会**定期往容量层去写**。

而全闪配置下，所有写操作先经过SSD，但是**并不会定期往容量层回写数据，只有当SSD上数据不再是热点数据(不再变化)时才会往容量层写**，由此提高容量层SSD寿命。

2.2.3 PCIE闪存盘 和 SSD的选择

> 关键词：性能需求，预算，容量

在计划使用PCIE闪存盘(nvme)时需要从价格、性能及容量三个方面考虑：

SSD 使用SATA接口，因此目前受限于SATA的6Gb/s接口(虽然SAS口可以达到12Gb/s，但是兼容SATA也就只有6Gb/s)。PCIE的闪存盘使用PCIE接口，接口类型为3.x时最大可达到32GB/s的带宽(PCIE闪存盘可以使用16个通道传输数据，每个通道单向有效带宽为1GB/s，双向2GB/s，更多详情请查看PCI-E百科，Wiki。

使用PCIE闪存盘也可以减少存储控制器的负载，因此有助于提高VSAN性能。

NVMe 在 VSAN6.1 开始受支持，NVMe相比SSD 延迟低，速度快，但是价格也更高。NVMe体积也一般比SSD大，容量可以做到更大，现在有6.4T的NVMe，而SSD才是4T（当然这一数据随着闪存发展会变化很快）。

PCIE 选择时也要注意服务器PCIE可用插槽位，当服务器磁盘组设计过多，加上NIC、HBA卡等，服务器自身的PCIE插槽可能不够用。

但如果使用SSD，SSD会占用HDD的槽位，如果服务器数量有限，但要求大容量，则可能选择nvme会好些。

2.2.4 闪存寿命

在VSAN6.0之后，使用TBW(TB字节写入)来计算寿命，而在此之前使用DWPD(每日整盘擦写的次数)来算寿命。

当使用TBW替代DWPD时，VMware允许**在容量层**使用DWPD不大，但是容量较大的盘替代DWPD较大，但是容量小的盘。

举个例子：一块200G的盘DWPD为10的盘，可以用一块400G，DWPD为5的盘代替，因为他们TBW是相等的。

而如果没有引入TBW的概念，VSAN规定SSD的DWPD要达到10，DWPD=5的盘则不会通过VSAN认证。

对于全闪存VSAN，VSAN要求缓存盘的TBW要达到4/每天，按照5年的使用寿命，总TBW为7300 (Intel数据中心级硬盘S3610可以达到10PBW)。容量层的SSD则可以根据需求使用寿命较低的盘。

2.2.5 快照开销

根据这篇文章，vSphere建议快照只保留24-72小时，最多32个快照，为了保证性能建议只留2-3个快照。

VSAN 6.0混合模式下，创建多个活动的快照可能很快将缓存耗尽，影响性能。VMware建议在快照使用频繁的场景下，将10%的缓存调整为15%。

2.3 磁盘(容量层)设计

> 关键词：类型、容量、数量、速度、价格

VSAN兼容列表里，有三种磁盘可以做容量层：SAS、NL-SAS、SATA盘。

按照最大容量对比：SATA 可达4T， NL-SAS 可达2T， SAS最大1.2T。

价格对比： SAS > NL-SAS >SATA

存储策略中有一项：**条带宽度**，如果对此项进行了设置，则必须要考虑磁盘的数量。条带宽度为1时，一个大小10G的object 会产生一个component，放在一个磁盘上。如果条带宽度为2，则这一个object可能被分成两个5G的component放在两个不同的磁盘上；

举一个极限的例子：三主机VSAN，每个主机一个磁盘组，每磁盘组一个SSD一个HDD，如果将条带宽度设置为2，则此虚拟机不能被创建，因为VSAN无法将虚拟机的一个对象拆成两个组件(组成逻辑上的RAID0)存放在**不同磁盘**中，磁盘数量不够。

VSAN最大支持64TB的虚拟机，如果环境中存在大虚拟机时，要考虑VSAN能不能存的下这个虚拟机，以一个三节点VSAN为例：剩余空间200T，按理是可以存下62T的虚拟机(只占用62*2=124T空间)，但是主机A剩余50T，主机B 50T，主机C 100T，这个VSAN环境是不能创建62T的虚拟机。

VSAN可以达到90%的读命中率，意味着有10%的读操作需要从HDD中去读取，所以HDD的数量多了，可以提高读操作的IO性能。

设计考虑：如果考虑到性能，建议优先采用10K以上的SAS盘。总容量不变的情况下，多个小容量HDD会比几个大容量组成的VSAN环境有更好的性能。一个环境**最好使用相同型号的HDD!**

2.3.1 磁盘总容量设计

> 关键词：30% 容量冗余

VSAN 总容量一般由需要的空间和FTT决定。当客户需要100T实际空间时，当FTT=1，那么裸磁盘总容量则为200T。

但是！还需要考虑到冗余，即当一台主机挂掉后，原来存在于这台主机上的数据需要能够在其他主机上恢复(rebuild或是resync)。

FTT=1时，就有很大风险，其他任意一个HDD或主机损坏，VSAN则会出现数据丢失。

VSAN 规定，当磁盘使用率达到80%的时候，会自动进行重平衡(Rebalance)操作，将磁盘上一部分数据转移到其他磁盘上，维持集群内磁盘使用率均衡的状态。VMware建议我们的总空间使用率比80%再低10%，也就是70%。*当然不一定这样严格，只是要知道总容量使用多于80%的时候，会有Rebalance操作，会对性能造成一定影响。*

所以按照开头的例子，实际裸容量要设计为285T左右。

2.3.2 文件系统格式开销

VSAN有自己的文件系统格式，到目前6.2使用v3版本，v3版本的开销为1%+duplication metadata，而Duplication metadata随着数据类型而变动很大。 VMware依然建议预留30%空间。

Virtual SAN version	Format Type	On-disk version	Overhead
5.5	VMFS-L	v1	750MB per disk
6.0	VMFS-L	v1	750MB per disk
6.0	VSAN-FS	v2	1% of physical disk capacity
6.2	VSAN-FS	V3	1% + deduplication metadata

2.4 I/O 控制器设计

> 关键词：每I/O controller支持的磁盘数、数量(考虑到单点故障及性能)、队列深度、直通模式

首先，所选择的I/O controller一定要在VSAN兼容列表内

VSAN支持多IO controller的ESXi主机，每个主机上最多35块硬盘(5个磁盘组，每个磁盘组7个hdd)。

有些控制器支持挂16个硬盘，一台主机安装两个这样的控制器能支持32个硬盘，几乎达到最大磁盘限制。

有些控制器只有8个接口，那么就需要4~5个控制器才能满足35个盘。

当主机有单个控制器，即使配置了多个磁盘组，单个控制器会引起单点故障。**多个控制器则可以避免这一点，同时也会提高性能。**

控制器队列深度，如果队列深度过小，在VSAN重同步数据的时候可能会影响性能。建议选择队列深度至少为256的，越大越好。

VSAN 支持**RAID0和直通硬盘**。推荐使用直通（JBOD）模式，使用RAID0时需要做很多额外的操作，例如virtual group创建，在更换硬盘时也需要相同的操作。

控制器最好没有缓存，因为VSAN有自己的缓存机制，所以无需在控制器上设置缓存，如果有些RAID卡必须设置缓存，那么请设置成100% read。

高级特性：有些控制器有高级特性，例如HP有个叫Smart Path的功能(LSI的fast path)用来做加速。VMware建议关掉这些高级功能。

三、存储策略

3.1 对象和组件

一个虚拟机由多个对象组成，例如一个VM主目录，vmdk文件，swap文件，快照文件。

在VSAN 5.5中，每个组件会产生 2M 的metadata，在VSAN6.0中，如果使用v2格式，每个组件占用4M空间。但是相比组件，这点开销很小，所以一般不予考虑。

在VSAN 5.5中，如果FTT>=1，则每个对象必须有一个仲裁(witness)，仲裁不存储数据，只存储元数据(metadata)，它在故障发生时裁决集群中是否存在满足FTT的数量的组件。5.5的仲裁规则是"**大于50%的组件**"，即要保证一个对象可用，可用组件数必须大于50%。

在VSAN 6.0中，决策方法变了。6.0中每个组件都含有一个**投票值**(a number of votes)，值>=1。然后，规则变为"**大于50%的投票**做裁决。因此会出现多个组件分布式存储，没有仲裁，但是依然能够保实现容错。 **但是，通常在6.0中很多对象还是会有仲裁的**。

3.2 存储策略中的设置

> 关键词: 条带数、内存预留、强制置备不要随意使用!

VSAN中有5个必须的策略(前五项)：

1.每个对象的磁盘带数(NumberOfDiskStripesPerObject)：

它定义一个VSAN对象的每个副本**最少**跨越多少个HDD进行存储。实际中VSAN的条带可能会多于策略所定义的数量(原因看前文，每组件大小限制)。

条带会有助于提升某些高 I/O 需求虚拟的性能，但是并不能确保提高性能。VMware建议普通场景下不做设置，保持默认的1，针对特殊性能需求的虚机，可以适量调整大，但是一定要考虑到FTT及主机上硬盘数。

2.闪存读缓存预留(FlashReadCacheReservation)：

之前我们提到建议缓存/容量比为1:10以上，这些缓存是平均地给所有虚拟机来使用。在VSAN中可以通过读缓存预留策略给一个或多个虚拟机预留缓存。(此策略仅对混合模式VSAN有效)

此预留策略是以虚拟机逻辑空间(也就是置备的空间)大小的百分比来定义的(因此如果设置100%预留，则缓存大小就是VM大小，会很浪费)，**请慎重使用此策略，仅在读性能明显很差的虚拟机上使用**。

风险举例：用户需要10台VM，每台瘦置备200G空间，但是实际占用100G。按照10%的缓存/容量比，用户需要购买100G闪存，其中70G作为读缓存。但是如果用户需要给每个虚拟机强加5%的读缓存预留，那么所有虚拟理论上会被分配到5%*200G*10=100G读缓存，明显会比正常设计的读缓存大。这种情况下反倒会影响虚拟机性能。

3.允许的故障数(number Of Failures To Tolerate)：

一个基本概念：FTT=n，则在VSAN中每个对象会有 n+1个副本，需要 2n+1 台主机才能满足这个策略。例如FTT=2，每个对象会有3个副本，至少需要5台主机组成的集群。

FTT 最大为3最小为0(但是生产环境不会用0)，当vmdk大于16T时，FTT只能为1。

VSAN 6.0中加入了故障域的概念，允许将几台相同属性的主机(例如使用同一个PDU，在同一个机柜)加入一个故障域中，允许他们同时故障。在此之前每台主机是一个故障域，而现在可以将一组主机视为一个故障域。所以当FTT=1时，至少需要3个故障域(每个故障域至少一台主机)，关于故障域后面有更详细的介绍。

4.强制置备(Force Provisioning)：

强制置备允许VSAN忽略FTT、条带宽度、读缓存预留等策略来创建虚拟机。

如果VSAN当前资源不满足需要创建虚拟机的策略，VSAN会使用极为简单的策略：FTT=0，条带宽度=1，读缓存预留=0 来创建这个虚拟机。也就是说这个虚拟机只有一份数据(对象空间预留策略也会被忽略)。

VSAN并不会尽最大程度创建尽可能满足policy的虚机，而是直接创建上述简单策略的虚机。例如FTT=2时，VSAN发现环境无法满足此策略，它不会去创建一个FTT=1的虚拟机，而是直接创建FTT=0的虚机。再如FTT=1，条带宽度=10，VSAN可以满足FTT=1，但是不能满足条带宽度=10，所以VSAN直接创建FTT=0，条带宽度=1的虚机。

当VSAN集群中的资源可以满足虚拟机的policy时，VSAN会**立即占用这些资源**去满足虚拟机的policy。

在6.0中，VSAN允许在主机进入维护模式(或者移除磁盘，磁盘组)时将数据**完整迁移(Full data evacuation)**。如果某个对象因为强制置备而成不合规的状态，那么完整迁移的操作等同于**保证数据可访问(ensure accessibility)，即允许对象减少可用性。**

5.对象空间预留(object space reservation)：

系统管理员始终要注意超额置备的问题。默认VSAN以瘦置备的方式创建虚拟机，　　对象空间预留(OSR)用于指定一个对象在逻辑预留多大空间(也就是类似于厚置备)。此策略值的范围在0%~100%之间，0%是默认的值，等同于瘦置备，100%可以视为等同于厚置备延迟置零。
VSAN有机制可以防止超额分配，例如当主机上可用存储空间不能满足FTT或是条带宽度要求，虚拟机创建会报错。

6.对象的IOPS限制(IOP Limit For Object)：

使用此策略可以限制一个对象或是一台虚拟机的IOPS占用。

有两种情况可以用此策略实现：限制一些非重要业务的IOPS，保证其他虚拟机的性能；实现同一资源池，分层的服务。

IOPS限制使用32KB的读写操作作为基准，即进行16KB的读写操作时，VSAN视为一个IO，当进行64KB的读写操作时，VSAN视为两个IO。

7.禁用对象校验(Disable Object Checksum)：

自VSAN 6.2之后，VSAN引入了对象检验功能，用于避免读写操作时由于硬件、软件(内存、驱动器等)导致的数据损坏。在VSAN 6.2(VSAN存储格式为v3)此功能默认开启的。每个读操作都会进行校验。同时每年scrubber会对所有一年内未读写的块进行校验。scrubber的校验周期可以改短，但是要注意这个操作是有额外后台开销的。

此功能会有额外的内存、CPU和存储开销，如果觉得没有意义可以启用此策略，关闭校验。

8.容错方法(Failure Tolerance Method)：

6.2 之前的VSAN版本只支持RAID1，6.2 全闪存模式下支持RAID5/6。

RAID 1有更好的性能，但是容量需求较高，RAID5/6在提供冗余的同时，相比RAID1更能节省空间。

当FTT=1时，使用RAID5/6(纠错码)大致占用1.33倍空间，使用RAID1则占用2倍空间。假如一个虚拟机为20G，使用RAID5/6实际占用27G，使用RAID1占用40G。

当FTT=2时，使用RAID5/6(纠错码)大致占用1.5倍空间，使用RAID1则占用3倍空间。假如一个虚拟机为20G，使用RAID5/6实际占用30G，使用RAID1占用60G。

和传统使用RAID5/6的存储一样，这样的容错方式会有额外的开销，但是VSAN仅在全闪配置下支持，SSD的性能很好所以做RAID5/6的开销可以忽略不计。

另外使用RAID5/6时，需要的主机数也有所不同。

FTT=1，做RAID5至少需要4台主机。

FTT=2，做RAID6至少需要6台主机(四台主机存放数据，两台做校验)。

3.3 虚拟机主页(VM namespace)及交换文件(Swap)考量

> 关键词：Swap文件为厚置备，可能占用很多空间。计算容量时需要考虑内存快照。

在VSAN中虚机以对象的形式保存，每个虚拟机都会有一个namespace。同时当虚拟机开机时，一个swap对象会被创建。namespace和swap这两个对象都不会继承存储策略，他们有自己的策略设置，会影响VSAN存储容量的设计。

VM namespace：VM namespace在VSAN中是个256G的瘦置备对象。因为一些策略(如对象空间预留，闪存读缓存预留)对于namespace没啥用，所以namespace可以忽略这些设定。默认namespace采用如下设置：

每个对象的磁盘带数：1

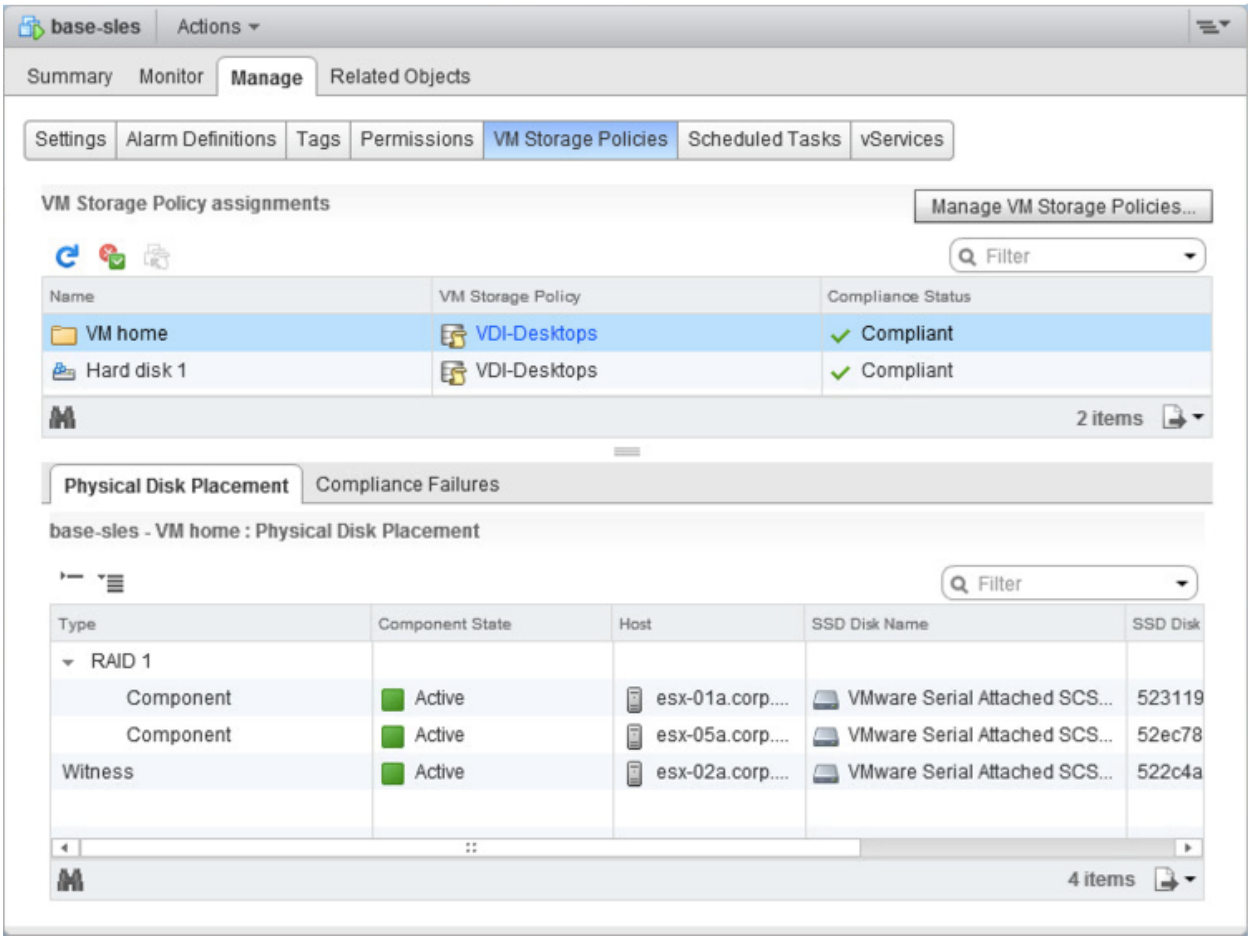
闪存读缓存预留：0%

FTT：继承VM的策略设置

强制置备：继承VM策略设置

对象空间预留：0%

下图是一个FTT=1时，虚拟机namespace的截图，可以看到它有一个RAID1，含有镜像的两个组件。还有一个保存在第三个主机的仲裁。



VM Swap

VM Swap也有自己的默认策略：

每个对象的磁盘带数：1

闪存读缓存预留：0%

FTT：1

强制置备：开启

对象空间预留：100%(厚置备)

在VSAN 6.2中有一个高级选项可以禁用VM Swap的厚制备，禁用后swap文件预留空间为0%，可以极大的节省空间（想象你环境中有500台虚拟机，每台4GB内存，如果都开机光Swap文件会占用4T空间）。需要在虚拟机关机后设置此选项。

```
esxcfg-advcfg -s 1 /VSAN/SwapThickProvisionDisabled
```

更多关于如何进行此高级设置，请看此文章。

3.4 快照

> 关键词：快照会占用空间，在快照使用较多的环境请注意容量预留

对VSAN上的虚拟机进行快照时，创建的Delta盘和母盘使用相同的policy。

如果对虚机进行内存快照，则会产生内存快照文件，在VSAN5.5时，内存快照保存在namespace中，因为namespace有255G的限制，所以需要内存小于255的虚拟机才能完成内存快照操作。在VSAN6.0后，内存单独拿出来成为一个对象保存。在做VSAN设计时需要考虑到内存快照的容量。

3.5 动态调整策略

VSAN支持动态调整策略，调整策略时**可能**会临时占用VSAN部分空间。如果系统资源不满足修改后的策略，reconfiguration会失败。

例如，将VM的FTT=1修改为FTT=2时，**VM原有的各对象和组件都不会变**，只是再增加一个副本。

但是如果修改了每对象磁盘带数，例如将1修改为2，很可能虚拟机对象的组件大小会发生变化，原来一个10G的组件可能拆分为两个5G的组件保存在不同的磁盘上。这时候，原虚拟机占用20G空间，调整策略时，临时会多占用20G空间。在策略应用完成后，原来的对象会被丢弃。

3.6 有可用空间不代表可以置备出虚拟机

VSAN的各种策略可能让虚拟机创建失败。例如磁盘不够时，每对象磁盘带数过大会导致置备失败。FTT结合每磁盘带数也可能导致置备失败。VSAN磁盘利用不均衡，导致实际可用磁盘数/主机数不能满足策略要求（有手动均衡磁盘利用的命令）。

四、主机设计

4.1 CPU 设计

每主机CPU数，CPU核数，VM的vCPU需求量，vCPU-CPU融合比，为VSAN预留10%的CPU资源。

4.2 内存设计

VM的内存需求量。当VSAN满配(5个磁盘组，每个磁盘组7块硬盘)时，需要至少32G内存。

4.3 启动设备设计

> 关键词：建议使用 SATADOM作为 启动盘

更多详细介绍见 这里 以及 这里。

VSAN 5.5开始支持从USB设备或者SD卡启动。

VSAN 6.0除了支持USB和SD，还支持SATADOM。

在将ESXi安装在USB和SD卡上后，ESXi的日志文件及VSAN traces(默认保存在scratch目录下，但此目录挂在RAM disk中)会保存在易失性存储RAM disk中，因此重启后日志就会丢失。

VSAN环境下也不建议将日志放在VSAN存储上，因为如果VSAN存储故障，则上面的日志也会丢失，为未来排错造成困难。

SATADOM 相比 SD 卡和 U 盘有更高的寿命和性能，所以日志可以保存在这些设备，但是对设备有些要求：容量>=16G，寿命： 512-1024 TBW *(VSAN6.1及之前) */ 384 TBW*（VSAN 6.2） 。

关于日志重定向及scratch文件夹重定向请见此KB1033696 。

4.4 尽量不要使用只提供计算资源的主机

因为VSAN有每主机最大组件数限制（VSAN5.5最大3000，6.0最大9000），所以可能置备的虚拟机数量会受到限制。

4.5 刀片服务器及外置磁盘柜支持

VSAN支持刀片服务器，但是受刀片服务器自身的限制，本地能够提供的容量有限。在VSAN 6.0之后开始支持外置磁盘柜，因此使用刀片服务器做VSAN变成一种可以落地的方案。如果机架式服务器本地硬盘有限，也可以外挂磁盘柜进行容量扩充。

VSAN 6.0支持的外置磁盘柜型号有限，如果需要使用此方案请务必检查VMware兼容性列表。

4.6 电源管理

> 关键词：建议关掉节能功能

电源管理会影响整体的性能，这个在虚拟化中一直被提及。当启用电源管理时，某些对处理性能延迟感知很明显的应用实际性能会不如预期。最佳时间建议将电源管理模式修改为'balanced'，避免服务器进入节能模式。 更多详细请看https://kb.vmware.com/kb/1018206

五、集群设计

5.1 2/3 节点设计

> 关键词：建议一套VSAN环境至少配置4节点

VSAN支持2节点及3节点配置，这样的配置和4节点及以上的表现会不一样。尤其是，当有主机宕机后，没有富余的资源重建丢失的组件，来保证FTT，在单台主机故障期间，容不得再有任何故障发生，否则数据会丢失。

另外使用2、3节点配置，如果有主机进入维护模式，无法使用“full data mirage”

所以，我们建议VSAN环境至少4台主机起，且一定要有足够多的资源预留。

5.2 vSphere HA

VSAN可以和vSphere HA结合使用。效果和使用传统存储时一样。

在网络隔离的情况下，vSphere HA有感知 VSAN 对象的功能，**但是这里要注意，vSphere HA与VSAN网络隔离并无绝对关系，VSAN网络隔离不会触发vSphere HA**（解决此问题的办法要么从物理角度避免VSAN网络中断，要么将管理网络和VSAN网络运行在同一个vDS中，或者修改HA的高级参数，将VSAN地址作为HA隔离检测地址）。

在**HA发生且出现网络隔离时(比如一个主机所有网络中断)**，如果一个 VM 原来所在的主机无法访问此VM需要的组件，它会被 HA 移动到可以访问此VM需要的组件的主机上，也就是说，HA能检测出来将虚拟机放在哪个分区合适，而不是随机去在任意一台主机上开机。

vSphere HA和VSAN搭配时有以下几点需求：

1、vSphere HA需要使用VSAN的网络进行通信

2、在开启VSAN之前需要关闭集群的HA，在VSAN创建完成后再打开集群的HA

3、vSphere HA不使用VSAN datastore作为存储心跳

VSAN 和 vSphere HA是解耦和的，一般开启vSphere HA时，vSphere会检查是否有足够的CPU和内存资源满足HA，使用VSAN时，HA不会去检查是否有足够的存储资源。当单个主机故障(组件是absent的情况下)，VSAN等待60分钟，60分钟后会在其他剩余存储资源上重建损坏主机上的组件，让虚拟机变成合规状态。

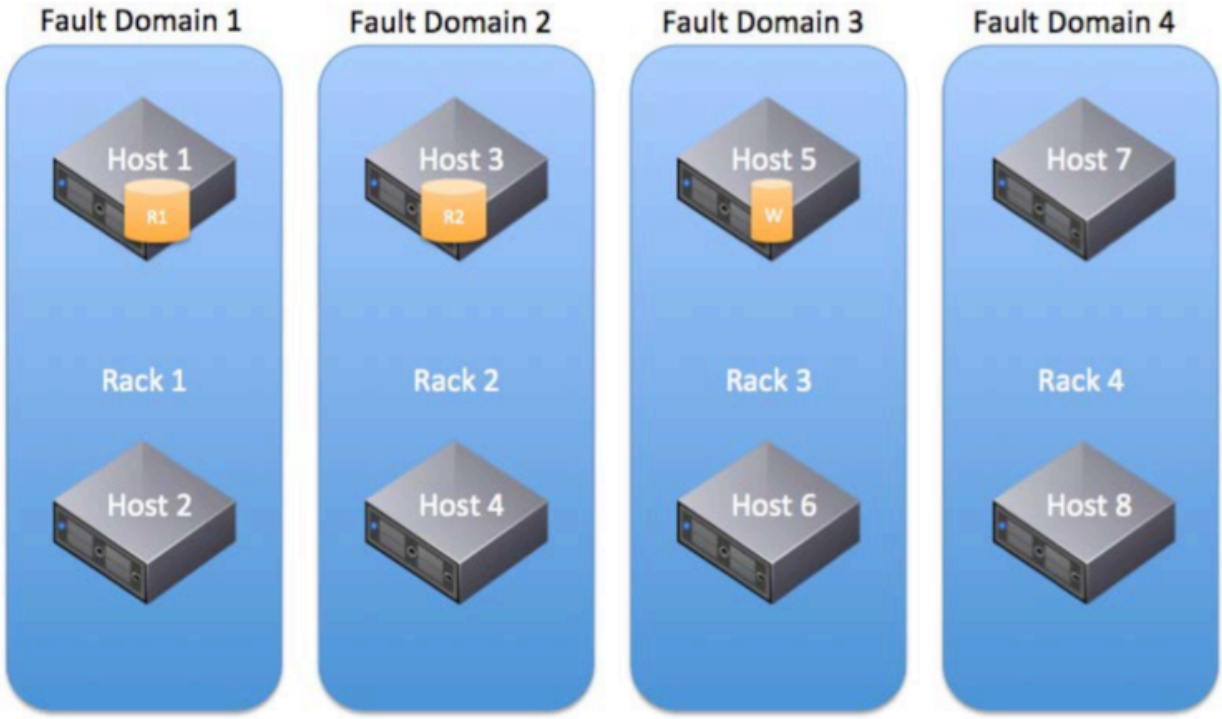
5.3 故障域

> 关键词：建议在较大环境(多个机柜)使用故障域

使用FTT来避免故障存在一种问题，就是当多个主机在同一个机架(通常会使用同一个PDU，同一个置顶交换机)时，整个机架故障会存在很大风险，如下图：

一个虚机很可能将数据以以下方式保存，所以当Rack1整个故障时，虚拟机失效组件>50，将会变成不可用状态。

如果引入故障域的概念，将一组主机作为一个故障点存在，那么VSAN则会避免将一个虚机的多个副本和仲裁保存在同一个故障域内。设置后虚机的数据会以下列方式存放。



在较大环境中，VMware推荐使用故障域避免单个机架故障。此时，同普通的设计一样，要考虑整个机架的故障后有可用资源接管。

5.4 去重和压缩设计

VSAN 全闪配置下支持去重和压缩。可以减少容量层磁盘空间的使用。它针对集群开启，只能在全闪配置下开启。

当启用此特性后，组件会平均分布在磁盘组中所有容量盘上。这样就避免了组件的重平衡，也无需为了均衡随机读性能而调整条带宽度。

当在开启去重和压缩后，使用对象空间预留策略，那么预留的空间不会被进行压缩及去重。

六、确定工作负载是否适合使用VSAN

> 关键词：在工作负载较重的场合，使用全闪存配置的VSAN

通常VSAN适用于大部分工作场合。

在使用混合配置时，需要注意应用程序如何使用缓存，有少部分应用对缓存的使用并不友好(没有固定热点数据、缓存命中率低)，例如：数据库的完整扫描，一个较大的数据库负载，内容很多的资源库，备份及恢复，等等类似的工作负载。

在遇到这些工作负载的时候，性能会取决于容量层磁盘的性能，例如有多少个可用的磁盘，他们速度如何，他们还承载了多少其他的工作负载。

相反，全闪配置总是可以提供很高的性能、较低的延迟，不会因工作负载而影响太多。

VSAN Ready Node 文档中介绍了Ready node服务器的一些参数，包括可以支撑的虚拟机数，能提供的IOPS等。

七、使用 View planner 进行VSAN规划

当 VSAN 用于 View 桌面环境时，可以使用View Planner 工具进行工作负载的模拟，进而得出较为真实合理的 VSAN 配置需求。

View planner可以模拟几种不同的工作场合(基础任务，办公，和高级用户)。进行模拟时，它会随机打开一些 Windows 常用的软件，模拟用户的操作，例如打开、保存、关闭、最小/最大化窗口；查看网页，编写文档、表格、ppt、观看视频、收发邮件、压缩文件等。View Planner使用专有的 watermark 技术量化用户的体验，测量应用延迟。

更多关于 View Planner 的信息请看这里。View Planner 的使用案例请看这里。

八、VMware infrastructure Planner - VIP

当前很多企业使用了 VMware 服务器虚拟化，更进一步，可能会发展成虚拟化、池化和自动化，即 SDDC。

VMware Infrastructure Planner 可以收集一个虚拟化环境中各种资源的使用量，告诉用户如果部署 vCloud 等 SDDC 产品后可以节省出多少资源。VIP提供直观的图表，能够展示不同资源的使用情况。

更多软件信息请看这里。

九、设计案例一

客户计划部署100台虚拟机，使用混合配置。

每台虚拟机8GB RAM、2vCPU、单个100G vmdk。

使用VSAN 6.0 v2格式。CPU融合比为5：1。

预计系统及应用会占用50%的存储资源，但是最终要提供100%的存储资源。

使用的存储策略为 FTT=1，其他都使用默认策略。

ESXi 装在SD卡上。

在此规划中，不计组件元数据及仲裁的开销，它们可以忽略不计。

CPU

主机数：>=3

总CPU数：300vCPU/5=40 核

考虑到VSAN自身占用10%的资源，总需要44核。

客户计划购买**两路，12核CPU**，按照三台主机算会有72核CPU，远远满足需求，且当一台服务器宕机后，其他两台有足够资源启动原主机上的VM。

内存

总内存需求：800G

平均每台需要300G内存。考虑到其中一台主机宕机后其他两台要接管，需要将每台主机内存增加到512G。

此处需要注意服务器能够安装这么多的内存。

磁盘容量

因为是三节点VSAN这样的最小配置，所以不需要考虑主机容量的冗余。

存储空间需求(不考虑FTT)：100GB*100=10TB

存储空间需求(考虑FTT)：100GB*100*2=20TB

存储空间需求+虚拟机swap文件（考虑FTT）：(10T+100*8*)2=21.6T

因为vmdk都是瘦置备的，所以在计算缓存层的容量时，使用瘦置备的容量来计算。

预计需要使用缓存的容量(不考虑FTT)：50% * 10TB = 5TB

缓存需求：5*10%=500G

预计快照需求：本例暂不考虑

前面提到至少要有30%的空间预留：

需要的总裸容量*70%=存储空间需求

即，需要的总裸容量= 存储空间需求/0.7 = 21.6/0.7 = 30.9TB

使用VSAN v2格式，文件系统的开销 = 1% * 总裸未格式化容量。

总结出公式：**总裸未格式化容量=需要的总裸容量+(总裸未格式化容量*1%，即开销)**

总裸未格式化容量=需要的总裸容量/99%

总裸未格式化容量=30.9/0.99=31.2TB

考虑到实际购买，则每服务器需要 10.5 T左右的容量，可以是10块1.2 T 10K的SAS盘。（此处粗略计算1.2T可用容量约1.1T）

磁盘组可以设计成两个，两个IO控制器。

两个SSD，每个100G。

组件数考虑

VSAN 6.0支持每主机最大9000个组件。

当前100台虚拟机，每个虚拟机至少包含如下组件：

1 X VM namespace

1 X VMDK

1 X swap

0 X snapshot

算上FTT=1，及仲裁：

2 X VM namespace + witness

2 X VMDK + witness

2 X swap + witness

0 X snapshot

总计100*9=900个组件，远小于9000/主机的限制。

十、设计案例二

客户计划部署400台虚拟机，使用混合配置。

每台虚拟机12GB RAM、1vCPU、100G 启动磁盘，200G数据盘。

使用VSAN 6.0 v2格式。CPU融合比为4：1。

预计系统及应用会占用75%的存储资源，但是最终要提供100%的存储资源。

使用的存储策略为 FTT=1，条带宽度=2，其他都使用默认策略。

ESXi 装在磁盘上。

在此规划中，不计组件元数据及仲裁的开销，它们可以忽略不计。

CPU

主机数：>=3

总CPU数：400vCPU/4=100 核

考虑到VSAN自身占用10%的资源，总需要110核。

客户计划购买**两路，12核CPU**，按照5台主机算会有120核CPU，刚好满足需求。

但是当一台服务器宕机后，其他服务器没有足够资源启动原主机上的VM，如果再增加一台主机即总6台则满足冗余需求。

内存

总内存需求：400*12=4.8T

总六台主机，平均每台需要800G内存。考虑到其中一台主机宕机后其他服务器要接管，需要将每台主机内存增加到1TB。

此处需要注意服务器能够安装这么多的内存。

磁盘容量

因为是三节点VSAN这样的最小配置，所以不需要考虑主机容量的冗余。

存储空间需求(不考虑FTT)：300GB*400=120TB

存储空间需求(考虑FTT)：120TB*2=240TB

存储空间需求+虚拟机swap文件（考虑FTT）：(10T+400*12)*2=249.6TB

因为vmdk都是瘦置备的，所以在计算缓存层的容量时，使用瘦置备的容量来计算。

预计需要使用缓存的容量(不考虑FTT)：75% * 120 TB = 90TB

缓存需求：90*10%=9TB

预计快照需求：每个虚拟机两个快照，但是快照增量不超过vm总容量的5%。

总快照占用：5% * 240T=12T

总裸容量：249.6+12=261.6T

前面提到至少要有30%的空间预留：

需要的总裸容量*70%=存储空间需求

即，需要的总裸容量= 存储空间需求/0.7 = 261.6/0.7 = 373.7 TB

使用VSAN v2格式，文件系统的开销 = 1% * 总裸未格式化容量。

总结出公式：**总裸未格式化容量=需要的总裸容量+(总裸未格式化容量*1%，即开销)**

总裸未格式化容量=需要的总裸容量/99%

总裸未格式化容量=373.7/0.99=377.5TB**

存储配置一：

6台主机，每台平均需要63T，1.5T 的SSD

我们可以选择使用4T的SATA盘，虽然速度慢一些，但是应该可以满足需求。每主机需要17块(考虑4T实际容量3.7T左右)。

因为每磁盘组7个磁盘的限制，至少需要三个磁盘组。

三个磁盘组意味着三块SSD，每个SSD 500G。为了未来考虑，可以选用容量更大一些的。

IO 控制器可以多选购一个。

共20块磁盘，因此可用磁盘插槽至少需要20（如果使用PCIE闪存，则至少需要17个插槽）。

ESXi 装在磁盘上，还需要额外的一个插槽，所以是至少21(或者18)。

但是，如果要考虑单个主机故障后重建组件的容量需求。就需要将主机数加到7台(这样考虑最简单了，实际可以主机数依然为6，提高每台上的存储容量)。

一开始计算CPU和内存是按照6台算的，主机数变成7台后相应的需要调整。CPU资源不好调整，可以适度减少每台主机上内存以节约成本。

存储配置二：

上种配置中，使用了7200rpm的SATA盘，性能可能满足不了业务需求，因此可以考虑购买1.2T 10k RPM的SAS盘。

总共需要315块磁盘，每个主机最大能有7*5个磁盘，提供42T的容量。

至少需要10台服务器(考虑1.2T实际可用容量1.1T)。

需要考虑IO控制器是否能接管所有硬盘。

此种大容量需求，也可以考虑使用外置磁盘柜。

10台主机共需要9T的闪存盘，每主机5个磁盘组，需要5*200G的闪存盘。

刚才算过每主机需要35个安装HDD，需要再加5个插槽安装SSD(如果使用PCIE闪存则只需要考虑PCIE槽位数是否足够)

ESXi 安装在磁盘上，需要外一个插槽，即 41个。

设计11台保证冗余。这时候CPU和内存就需要重新进行设计，将原来6台的总量均分到11台服务器上。

CPU每主机可以选购8核的，内存每主机640G足够。

存储配置三-全闪配置(这部分原文很混乱，没考虑swap，且有错误)：

上述配置中，为了满足性能使用SAS盘，但受容量限制导致需要更多服务器来承载。第三种配置我们使用全闪配置，看看使用RAID5/6以及去重压缩能节省多少容量。

预计的启动盘去重压缩率为4x，因为虚拟机会使用链接克隆技术创建。数据盘压缩率2x。

单个启动盘的容量需求(FTT=1，RAID5，去重压缩率4x)：

100G*1.33/4=33.25GB

单个数据盘的容量需求(FTT=1，RAID5，去重压缩率2x)：

200G*1.33/2=133GB

Swap的容量需求(FTT=1，RAID5，去重压缩率1x)：

12*1.33/1=15.96GB

总裸容量需求（不含swap）：（33.25GB+133GB）*400=66.5 TB

总裸容量需求（含swap）：66.5+15.96*400=72.9 TB

预计快照需求：每个虚拟机两个快照，但是快照增量不超过vm总容量的5%。

总快照占用：5% * 66.5T=3.325 T

总裸容量：72.9+3.325 = 76.3 T

在全闪配置时也是按照vmdk占用容量计算。

预计需要使用缓存的容量：75% * 120 TB = 90TB

缓存需求：90*10% = 9TB

前面提到至少要有30%的空间预留：

需要的总裸容量*70%=存储空间需求

即，需要的总裸容量= 存储空间需求/0.7 = 76.3/0.7 = 109 TB

使用VSAN v2格式，文件系统的开销 = 1% * 总裸未格式化容量。

总结出公式：**总裸未格式化容量=需要的总裸容量+(总裸未格式化容量*1%，即开销)**

总裸未格式化容量=需要的总裸容量/99%

总裸未格式化容量 = 109/0.99 = 110 TB

按照CPU和内存计算需要6台服务器，但是呢，要保证数据冗余，此处CPU和内存已经是冗余状态，再增加一台主机显得多余，所以可以将110T平均到5台服务器上，每台提供22T，然后第六台增加22T的容量，这样总共132T空间。一台主机故障后，CPU、内存。

22T/主机的容量，SSD 容量范围也比较大，可以选择20块1.2T的，15块1.6T的等等(考虑到硬盘实际容量小一些)。

缓存/容量比为1:10，但是全闪配置下每个磁盘组又有600G的容量限制，所以以1.2T为例，每主机需要4个磁盘组。共24个磁盘组，每磁盘组400G SSD做缓存，共9.6TB满足计算的9TB。

最后检查组件数

当前400台虚拟机，每个虚拟机至少包含如下组件：

1 X VM namespace

2 X VMDK

1 X swap

2 X snapshot

算上FTT=1，条带宽度=2，及仲裁：

2 X VM namespace + witness =3

（2X2 X VMDK + 3 witness ）*2 = 14

2 X swap + witness=3

（2X2 X snapshot + 3 witness ）*2 = 14

总计400*34=13600个组件，按6台主机每主机约2267个，满足9000的大小上限。

即使当一台主机宕机，剩余5台时，每主机2720也满足。

对比一下三种配置方案：

使用SATA时性能不好，但是需要的主机可以比较少，唯一需要注意的是每台服务器是否能安装1T内存

使用SAS盘时，会需要很多主机来安装硬盘，从经济角度考虑可能并不推荐。此方案可以考虑使用外置磁盘柜扩充。

使用全闪配置时，容量需求会小很多，且SSD单盘容量可以很大，因此能非常容易地满足需求。

十一、下面是一些网站链接：

VMware Ready Nodes

http://www.vmware.com/resources/compatibility/search.php?deviceCategory=vsan

VMware 兼容性指南

http://vmwa.re/vsanhclc

vSphere 社区中的VSAN板块

https://communities.vmware.com/community/vmtn/vsan

博客

http://cormachogan.com/vsan/

http://www.yellow-bricks.com/virtual-san/

http://www.virtuallyghetto.com/category/vsan

http://www.punchingclouds.com/tag/vsan/

http://blogs.vmware.com/vsphere/storage

http://www.thenicholson.com/vsan

文档库

http://www.vmware.com/products/virtual-san/resources.html

https://www.vmware.com/support/virtual-san

VMware 支持

https://my.vmware.com/web/vmware/login

http://kb.vmware.com/kb/2006985 - 如何获得帮助

http://kb.vmware.com/kb/1021806 - VMware 产品位置

日志文件

http://kb.vmware.com/kb/2032076 - ESXi 5.x 日志文件

http://kb.vmware.com/kb/2072796 - 收集 Virtual SAN 支持日志

更多阅读

http://blogs.vmware.com/vsphere/files/2014/09/vsan-sql- dvdstore-perf.pdf - Microsoft SQL Server 性能学习

http://www.vmware.com/files/pdf/products/vsan/VMW-TMD-Virt-SAN-Dsn-Szing-Guid-Horizon-View.pdf - Horizon View VDI 设计指南

http://www.vmware.com/files/pdf/products/vsan/VMware-Virtual-SAN-Network-Design-Guide.pdf - VSAN网络设计指南