高可用存储网络设计

简介

在今天的企业环境中，高可用性不再是一个可有可无的考虑因素。随着数据增长速度的不断加快，数据可用性变得比以往任何时候都更加重要。随着企业和应用的不断发展，发展相关的数据中心基础设施的能力也变得非常关键。在互联网的推动下，经济的全球化趋势将企业的经营模式从 8×5 变成了 24×7。在这种“永不中断”的环境中，人们对高可用性提出了更加严格的要求。为了保持一个企业的正常运行，数据——企业最重要的资产——必须在任何时候都可供使用。不仅数据丢失会造成灾难性的后果，数据无法访问也会造成同样严重的损失。

尽管 99%的正常运行时间似乎已经是相当出色的成就，但是这种“高度可用”的环境每年仍然会停机超过 83 小时。这可能会给任何规模的企业造成严重的影响。在设计一个高度可用的解决方案时，必须考虑停机成本。拥有 99%正常运行时间的环境将导致一个金融经纪企业每年损失相当于 5.4 亿美元的收入和生产率。

	每小时成本
金融经纪	650 万美元
信用卡授权	260 万美元
家庭购物	10 万美元
目录销售	9 万美元
机票预订	9 万美元
远程订票	7 万美元

可用性	停机时间（分钟数/每年）
99.999%	5
99.99%	50
99.9%	500
99%	5000
90%	50000

数据来源：光纤通道行业协会[1]；Horison 公司

通过将正常运行时间提高到 99.999%，企业每年可以节约 54 万美元。

1. “在灾难发生时保障业务的连续性”；光纤通道行业协会；http://www.fibrechannel.com/technology/index.master.html

实现 99.999%并不是一项轻松的任务。但是，一个高度可用的存储基础设施是提高数据可用性的关键。它主要包含下列组件：低成本磁盘冗余阵列（RAID）技术，数据在一个群集系统中的多份复本，远程群集，存储网络（SAN），以及可靠的磁带备份。其中，SAN 架构可以实现覆盖整个企业的高可用性配置，它可以随着企业的发展而扩展，并保护您的数据存储投资。在设计一个高度可用的 SAN 时需要考虑很多重要的因素。

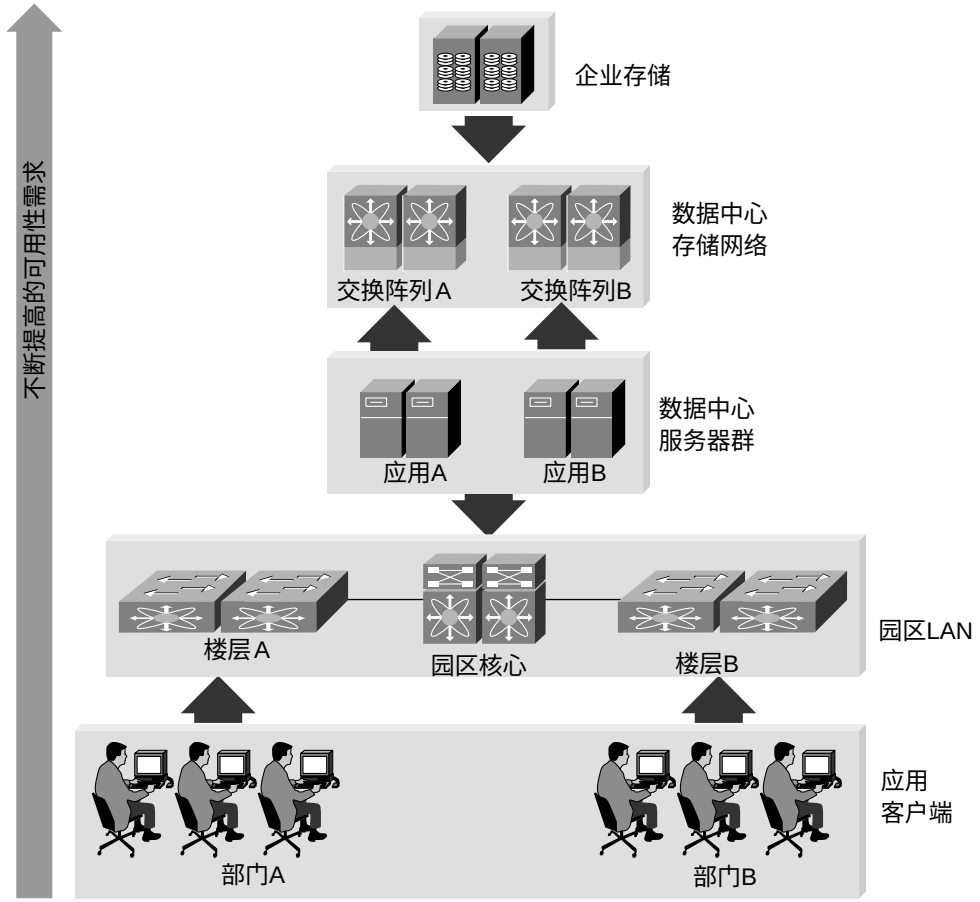
- 人为错误可能造成的影响和防止人为错误的方法
- 环境问题（例如电源中断、空调故障、管道故障）
- 基础设施设备的软件故障（交换机等）
- 计划内停机（软件升级、硬件维护）
- 黑客构成的威胁
- 基础设施设备的硬件故障（交换机等）

通过一个可靠的部署设计可以有效避免其中的一些事件，例如硬件故障和电源中断。但是其他一些因素（例如人为错误）并不能简单地通过设计消除。

存储系统的正常运行对于整个企业来说非常重要。每个员工都需要通过访问存储来制定关键的业务决策——无论是通过某个应用服务器还是直接从他们的工作站进行。当存储的可用性出现问题时，整个企业的运作都会受到影响。

为了避免出现这样的情况，必须确保最大限度的正常运行时间，以限制和消除任何可能的业务影响。

图 1 企业的高可用性优先等级



设计高可用性解决方案

在设计一个高度可用的存储环境时，必须使用一种端到端的方法。只考虑存储解决方案的各个组件并不够，还必须考虑下列因素：

存储子系统

存储子系统的三个方面对于设计一个高度可用的解决方案非常关键。

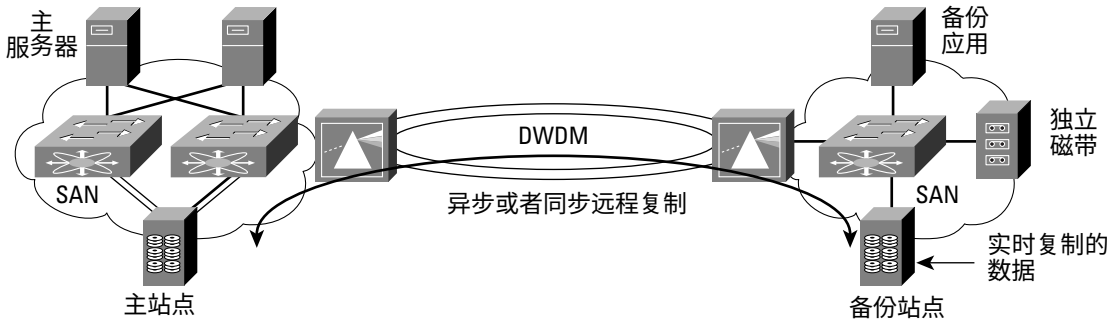
数据保护

- **冗余缓存**——几乎所有的存储子系统都采用了某种前端缓存。这种缓存可以通过将需要写入的数据放入缓存提高子系统的响应速度，如果直接写入磁盘，可能会导致很长的延时。当某个应用服务器发出一个写入命令时，存储子系统将把数据写到缓存中，这样可以大大缩短延时，继而通知应用服务器写入操作已经完成。这些数据将在稍后的某个时间写入磁盘，或者存入低速的后备存储器。很多子系统会为前端缓存建立镜像，以提高可用性。如果某个缓存发生故障，通过镜像的复本可以保存数据。
- **RAID**——几乎所有的子系统都利用低成本磁盘冗余阵列（RAID）提供更高的数据可用性，以保护数

据和提高存取速度。RAID 部署可能只是简单的 RAID1（将数据镜像到两个或者更多的磁盘）或者可以通过数据奇偶性计算而进行提前分段的 RAID 5。RAID 1 和 RAID 5 技术可以提供不同等级的保护和性能，但是两者都可以在磁盘发生故障时提供足够的可用性。

- **数据复制**——采用存储复制通常是为了防止整个子系统发生故障。尽管复制通常是以异步方式在较远的距离上进行，但是这种方式超出了本文的讨论范围。同步数据复制可以用于确保某个本地数据中心的数据可用性，同时可以最大限度地降低对重要应用的性能影响。这种复制可以通过存储子系统完成，也可以通过某个基于主机上的外部应用完成。在任何一种情况下，最终结果都是两个独立的存储子系统，每个子系统具有同一份数据的实时复本。

图 2 同步数据复制模式

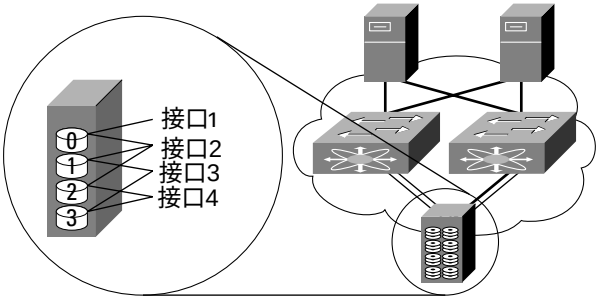


子系统连接

存储连接和存储本身的完整性一样重要。如果某个应用不能访问它的存储，它将不得不停止工作。因此，存储在存储子系统中的供应方式对于整个存储高可用性解决方案来说非常重要。

- **冗余接口**——连接必须是冗余的，以获得真正的高可用性。一个磁盘逻辑设备必须通过多个接口连接到存储子系统。这不仅可以在主机级别实现多路径，还可以通过磁盘子系统本身提供的两个物理连接实现更高的冗余性。

图 3 用于实现高可用性的冗余磁盘子系统接口



子系统硬件冗余

- **电源冗余**——电源对于存储子系统非常关键。大多数存储子系统都将双电源作为标准配置。此外，大部分具有前端缓存的子系统都为缓存提供了一定级别的备用电池。一些子系统使用容量较小的电池，只够为缓存提供几天的电能。也有一些子系统使用容量较大的电池，让整个系统可以运行足够长的时间，以便将数据从缓存转移到物理磁盘。
- **热磁盘备用**——大部分存储子系统都可以提供备用的物理磁盘。这些备用磁盘（每个子系统的备用磁盘的容量可能各不相同）只在某个磁盘表现出故障迹象或者突然失效时使用。子系统会监控各个物理硬盘，以发现潜在的故障迹象。如果子系统发现了故障迹象，发生故障的磁盘中的数据就会被复制到热备用磁盘。而且，因为 RAID 通常用于存储子系统，所以如果某个 RAID 群组的一个磁盘突然发生故障，一个热备用磁盘可以用于恢复丢失的数据。在任何一种情况下，子系统都能够恢复和访问未损坏的数据。

存储网络

可以在主机和存储之间提供连接的网络或者交换阵列也是整个高可用性解决方案的一个重要组成部分。最佳的设计实践可以被用于确保设计中不存在单点故障。这种设计实践还可以确保正确的冗余等级，因为过度的冗余可能会导致故障恢复速度的降低。

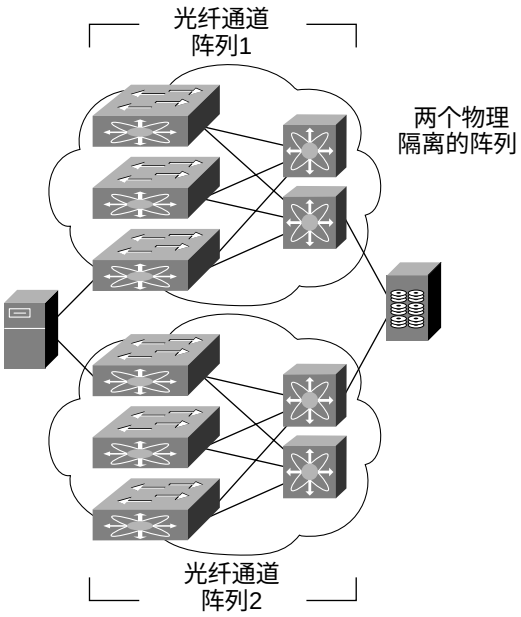
存储网络硬件

- **交换机硬件**——与存储解决方案中的所有其他硬件组件一样，光纤通道交换机的硬件必须是冗余的。在交换机级别的产品中，硬件冗余通常只限于双电源。这可以解决电源中断问题，但是不能解决其他交换机组件的故障。控制器(Director)级别的光纤通道交换机为存储网络带来了新的可用性等级。它们不仅可以支持冗余电源，而且每个其他的重要组件都是冗余的。控制模块可以提供故障转移功能。因此，控制器级的硬件有助于在系统中实现真正的 99.999%正常运行时间。

存储网络设计

- **交换阵列冗余**——光纤通道 SAN 中的另外一个需要关注的领域是交换阵列本身。每个连接到同一个物理基础设施的设备都处于同一个光纤通道阵列中。这使得 SAN 很容易受到交换阵列级事件的影响，这种故障可能会导致网络上的所有设备发生中断。各种改变（例如添加交换机或者改变分区配置）可能会影响整个交换阵列。因此，设计单独连接的交换阵列有助于限制这种事件的影响范围。思科的虚拟 SAN（VSAN）功能可以利用相同的物理基础设施提供一种复制这种环境的方法，即隔离事件。本文稍后将详细地介绍 VSAN 功能。

图 4 利用独立的交换阵列设计 SAN



- **交换机间连接（ISL）**——随着 SAN 的不断发展，交换机间的连接将变得更加重要。依靠交换机之间的单一物理连接会导致设计的整体冗余性的降低。如果某个连接发生故障，冗余 ISL 可以提供故障转移功能。

应用主机

主机总线适配器（HBA）是应用服务器和 SAN 之间的接口。与网络接口类似，它们被插入到服务器的总线插槽中。尽管大部分服务器所生成的输入/输出（I/O）都不足以超出一个光纤通道连接的承受程度，但是高可用性（HA）环境仍然必须具有两个 HBA。两个或者多个 HBA 可以提供多条访问存储的路径。这不仅可以在一个 HBA 发生故障时实现故障转移，还可以在 HBA 之间提供负载平衡。这种“多路径”可以通过多种方法实现。用户可以选择下列选项，为 HBA 提供高可用性：

- **子系统软件**——大部分主流存储子系统供应商都开发了多路径软件，提供经过认证的 HBA 的负载平衡和故障转移功能。一个典型的例子是 EMC 的 PowePath。这些软件通常都专门针对供应商的子系统设计，或者在使用供应商自己的子系统时可以提供某种增强模式。
- **卷管理软件**——一些基于主机的卷管理应用可以支持多路径。一个典型的例子是 Veritas 提供的动态多路径软件（DMP）。这种解决方案不是针对某个特定的子系统供应商的。
- **HBA 驱动程序**——一些 HBA 供应商目前可以在主机上的 HBA 驱动程序中提供多路径功能。这种解决方案并不针对某个特定的存储供应商，但是它会将多路径功能限制于某个特定的 HBA 供应商，并可能采用某种特殊的 HBA 模式。
- **操作系统**——很多操作系统（OS）现在就可以支持多路径功能。这使得多路径功能可以脱离存储子系统和 HBA。

提高存储网络的可用性

Cisco MDS 9500 系列多层控制器(Director)可以提供大量的硬件和软件功能，能够在光纤通道网络内部实现增强的可用性。

硬件功能

下面几节将介绍 Cisco MDS 9500 系列多层控制器的高可用性在硬件方面的表现。

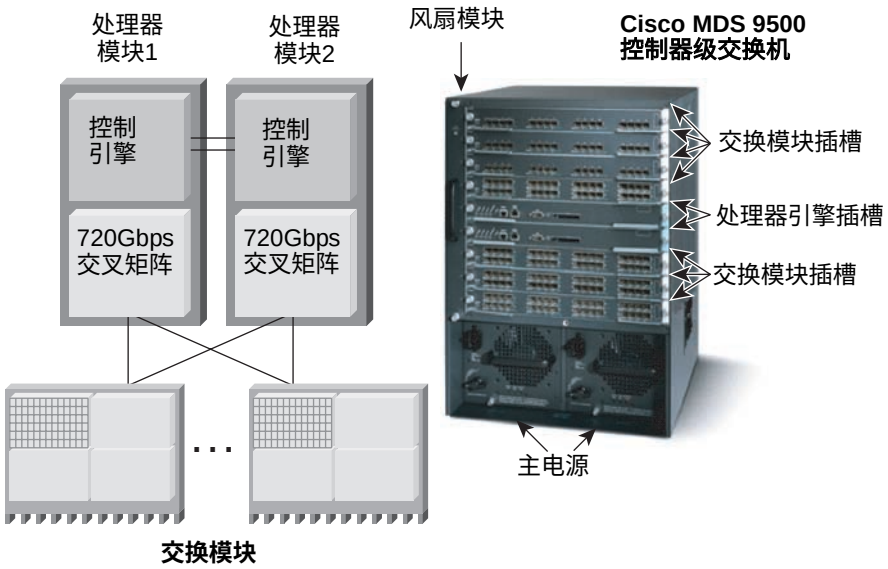
处理器模块

Cisco MDS 9500 系列多层控制器允许在机箱中安装两个处理器模块，以实现冗余性。每个处理器模块包含一个控制引擎和交叉矩阵(Crossbar)。控制引擎是核心处理器，负责管理整个系统。此外，控制引擎会参与所有网络控制协议的工作，包括光纤通道服务。在一个冗余的系统中，两个控制引擎以一种主/备用模式工作，从而确保始终有一个控制引擎处于工作状态。处于备用模式的控制引擎实际上处于一种全状态待命模式，这样它可以与主控制引擎支持的所有管理和控制协议保持同步。尽管备用的控制引擎并不会主动管理交换机，但是它可以不断地从主控制引擎接收信息。这使得交换机的状态可以保存在两个控制引擎之间。如果主控制引擎发生故障，备用控制引擎将会无缝地接替它的职能。

交叉矩阵(Crossbar)是系统的交换引擎。交叉矩阵可以在系统的所有端口之间提供一个高速的交换路径矩阵。每个处理器模块中都会嵌入一个交叉矩阵。因此，在拥有两个处理器模块的冗余系统中将会有两个交叉矩阵。但是，每个交叉矩阵的总交换容量为 720Gbps，可以为每个插槽提供 80Gbps 的带宽。因为 Cisco MDS 9500 系列的每个交换模块所使用的带宽都不会超过交叉矩阵提供的 80Gbps，所以系统能够以全速运行，即使只使用一个处理器模块。因此，在一个配置齐全的 Cisco MDS 9500 系列控制器中，系统不会因为一个处理器模块的插拔或者故障而发生任何中断或者损失任何性能。

处理器模块是一个可以热插拔的模块。在一个双处理器模块的系统中，这使得用户可以在不导致系统其他部分中断的情况下，移除或者更换该模块。

图 5 Cisco MDS 9500 系列交换系统



电源

Cisco MDS 9500 系列多层控制器可以支持双电源。电源以一种主－主配置运行，但是彼此独立。如果某个电源发生故障，单个电源足以为整个系统供电。每个电源都可以热插拔。每个电源都针对为整个系统供电而设计，从而可以更换发生故障的电源。

系统风扇

Cisco MDS 9500 系列多层控制器使用了一个风扇盘来冷却整个系统。尽管这看上去是一种非冗余的组件，但是风扇采用了一种 N+1 的冗余配置。风扇盘上的每个风扇都将受到单独的监控。如果某个风扇发生故障，系统将会通知终端用户。但是，系统可以在多个风扇发生故障的情况下继续正常工作。在正常的工作环境中，在四个风扇发生故障时，系统都不会受到任何影响。整个风扇盘可以热插拔。系统可以在不安装风扇盘的情况下工作长达 30 分钟。这让用户可以在系统正常工作时更换风扇盘。

软件功能

尽管传统的光纤通道交换机只依靠硬件冗余实现高可用性，但是 Cisco MDS 9500 系列可以提供一组强大的软件功能，从而可以在典型的存储网络中提高基于硬件的冗余性。

不中断软件升级

计划内的停机时间在每年设备停机时间中占很大的比例。 计划内停机的一个重要原因是升级网络设备的软件。这种升级可能是修复软件缺陷或者添加软件功能。无论什么原因，计划内停机都有可能会影响正常的业务。任何控制器级光纤通道交换机的一个重要功能是在不中断 SAN 中的流量的情况下在交换机上载入和启用新软件。

Cisco MDS 9500 系列多层控制器支持在正常工作期间升级处理器模块和交换端口模块的软件。在升级期间，用户可以选择升级两个处理器还是只升级主处理器。这让备用处理器可以在新版本运行的同时保持旧版本的程序。如果新版本的程序出现问题，就可以转而使用运行旧版本软件的备用处理器。这可以为软件升级提供最大限度的灵活性，同时为恢复到已知的、稳定的软件提供一条路径。

内部流程重启

Cisco MDS 9500 系列的一个非常独特的功能是可以重启一个出错的软件流程。处理器模块可以不断地监控所有的软件流程。如果某个流程出错，处理器可以在不中断交换机流量传输的情况下重启流程。这项功能有助于提高可靠性，因为如果流程可以重启，可能不需要对处理器进行故障转移。如果流程不能重启，或者继续出错，那么主处理器模块将可以将任务移交给备用处理器模块。

VSAN

目前，很多 SAN 设计人员都出于各种原因在建设独立的存储网络。在这种情况下，一个独立的存储网络指得是一个在物理上完全隔离的、用于将主机连接到存储的交换机或者交换机群组。下面列出了一些常见的原因：

1. **高可用性**——一种常见的做法是建设多个并行的交换阵列，并在并行的、物理上隔离的交换阵列中加入“ 多穴 ”主机和磁盘。一般而言，采用这种隔离方式的主要原因是确保各个阵列的服务（例如名称服务）彼此隔离。如果某个交换阵列服务发生故障，它就不会影响其他的并行交换阵列。因此，并行交换阵列可以提供从主机到磁盘的隔离路径。
2. **应用和备份交换阵列**——很多客户为他们的存储网络环境构建了至少两个物理上独立的交换阵列。他们的主要目的是将一个交换阵列专门用作应用主机，另外一个专门用于备份环境。利用这种方式，备份流量在物理上与主应用流量完全隔离。
3. **部门交换阵列**——很多客户选择为部门应用扩展独立的存储网络环境。在这种情况下，客户会为每个部门应用建立一个独立的、规模较小的交换阵列。
4. **多种 OS 的交换阵列**——有些客户的做法是为使用不同操作系统的主机建立独立的交换阵列。由于一些操作系统的特性以及它们发现和使用存储的方法，很多客户会用不同的阵列隔离环境。一个典型的例子是一个 Sun Solaris 阵列和一个 Windows NT/2K 阵列。

尽管每个原因都为建立独立的交换阵列提供了有力的依据，但是这种做法的确相当浪费。添加独立的交换阵列意味着购买更多的硬件、投入更多的资金，而硬件通常无法得到充分的利用。

为了实现同样隔离的环境，同时避免为了建立物理上隔离的交换阵列而花费大量资金，思科在 Cisco MDS 9000 系列中推出了虚拟 SAN（即 VSAN）功能。VSAN 可以提供在同一套基础设施上建立独立的虚拟交换阵列的能力。利用 ISL 链路上的一种基于硬件的帧标签机制，各个虚拟交换阵列都彼此隔离。EISL 链路

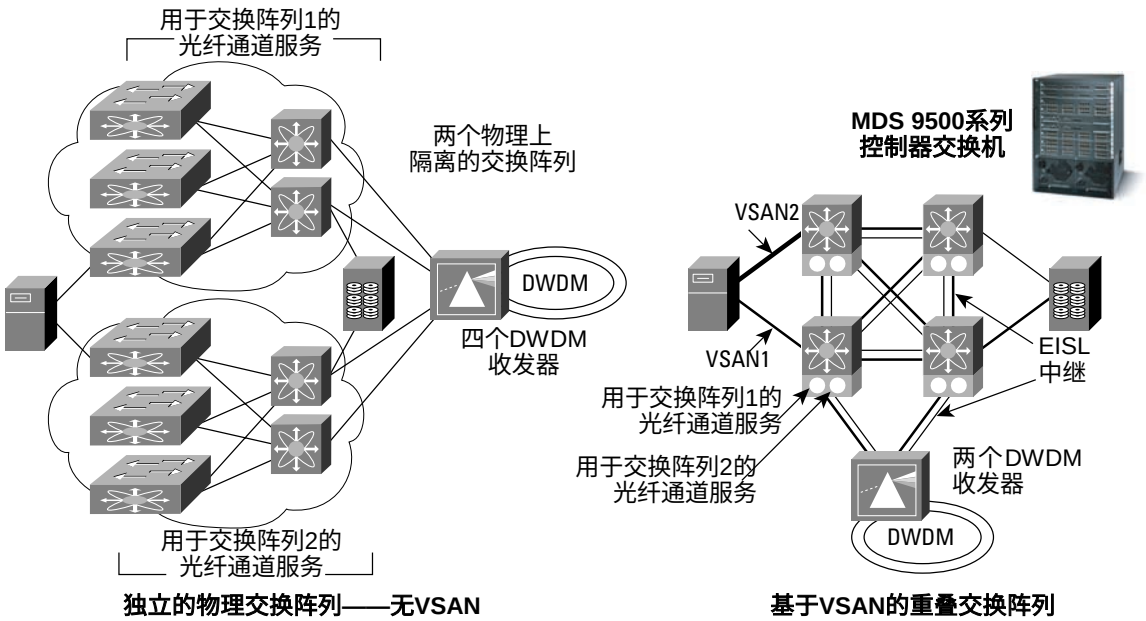
是一种增强的 ISL 链路，包括用于每个帧的附加标签信息，并可以支持互联任何 Cisco MDS 9000 系列交换机产品的链路。VSAN 的成员资格建立在物理端口的基础上，没有哪个物理端口可以属于两个以上 VSAN。因此，连接到一个物理端口的任何节点都将成为该端口的 VSAN 的成员。

VSAN 为用户提供了巨大的灵活性。例如，Cisco MDS 9000 系列产品可以在每个物理基础设施中支持 1024 个 VSAN。用户可以有选择地在一个 EISL（增强 ISL）链路上添加或者去除 VSAN，以控制 VSAN 的范围。此外，它还可以提供了特殊的流量计数器，以跟踪每个 VSAN 的统计数据。

VSAN 最受欢迎的特性很可能是它的高可用性。VSAN 不仅可以提供严格的硬件隔离，还为每个新的 VSAN 创建了一套完整的光纤通道服务。因此，当用户创建某个新的 VSAN 时，一套完全独立的服务（包括名称服务器、分区服务器、域控制器、别名服务器和登陆服务器）将会在支持这个新的 VSAN 的交换机上创建并启用。这种复制的服务可以在同一个物理基础设施上建设独立的环境，消除人们在 HA 方面的担忧。例如，在 VSAN1 中安装一个主分区集不会对 VSAN2 中的交换阵列造成任何影响。

VSAN 还可以通过一个远程基础设施连接远程数据中心的独立交换阵列。因为帧标签功能由硬件完成，并且包含在每个 EISL 帧之中，所以它可以通过密集波分复用（DWDM）或者粗波分复用（CWDM）等传输方式进行传输。因此，来自于多个 VSAN 的流量可以复合到一对光纤上，并可以在更长的距离上传输，同时仍然保持完全的隔离。VSAN 使用户可以利用一个通用的冗余物理基础设施来建设灵活的、隔离的交换阵列，以实现高可用性目标，从而将可扩展性提高到了一个新的等级。

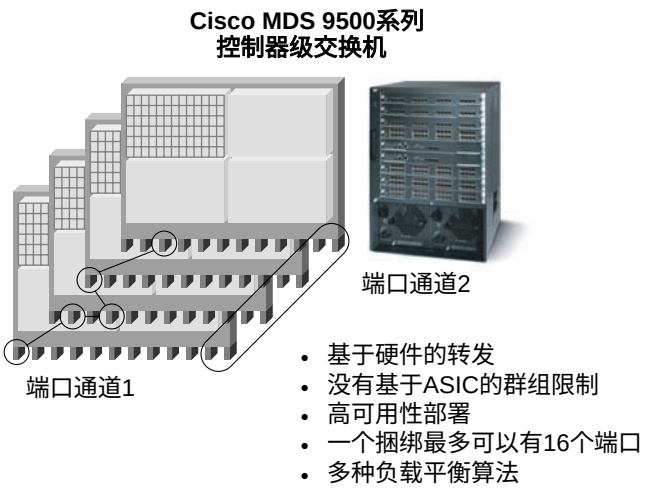
图 6 利用 VSAN 降低 SAN 的复杂性



ISL 端口通道 (Port Channel)

随着光纤通道矩阵的日益增大，用户需要更多的交换机来满足他们对于端口数的要求。ISL 可以实现交换机间的连接。域 SAN 的所有其他连接一样，这些连接必须是冗余的。利用思科的端口通道技术，用户最多可以将 16 个独立的物理连接整合到一起，在交换机之间创建一个逻辑 ISL。这不仅可以提供一个具有很高弹性的逻辑连接，还可以在两台交换机之间提供 32Gbps 的带宽。思科的端口通道技术的一个关键优势是让捆绑的网络连接可以位于交换机的任何一个交换模块的任何一个端口上。通过将物理连接拓展到多个交换模块，不仅可以防止连接故障（例如电缆中断或者光纤故障），还可以防止交换机模块发生故障。

图 7 Cisco MDS 9500 系列中的端口通道



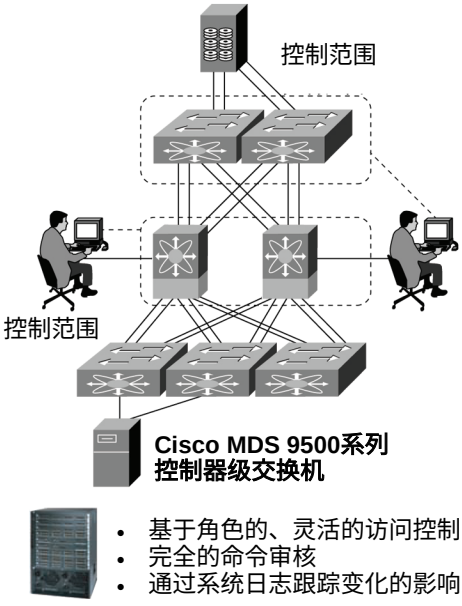
Cisco MDS 9500 系列控制器级交换机可以在端口通道中采用两种不同的负载平衡算法。第一种算法在帧进入端口通道之前检查帧的源和目的地 FC_ID。它会在硬件基础上，对帧的源和目的地 FC_ID 进行散列，并将其作为该流量在虚拟连接中应当采用的物理连接的索引。来自于该源-目的地对的流量将始终通过同一条连接传输。源-目的地 FC_ID 的不同组合将进行独立的连接决策，不一定通过同一条连接传输。从目的地到源的流量并不一定通过同一条物理连接传输，因为目的地端的交换机是独立地制定连接流量决策的。

Cisco MDS 9500 系列中的第二种算法是基于源-目的地 FC_ID 和操作的 Exchange_ID（OX_ID、RX_ID）的负载平衡。每进行一次操作，就会使用一个新的 Exchange_ID，并制定一个新的物理连接决策。这可以最大限度地提高整个端口通道的效率，即使在同一个源和目的地对之间。利用这种算法，同一对源和目的地之间的数据交换可以分布在端口通道的多个连接上，但是可以保证所有帧都按顺序处于任何一个特定的交换流程中。

基于角色的安全性

安全性通常与高可用性没有直接的联系。但是，导致停机的一个重要原因就是人为错误。一个用户可能会在不了解某个命令的效果的情况下错误地执行该命令。Cisco MDS 9500 系列多层控制器和矩阵交换机支持一种基于角色的安全机制，可以确保只有经过授权的个人才能访问矩阵中的关键功能。每个用户都会被分配一个角色（即组 ID），该角色将获得矩阵的某种特定的访问权限。这种访问权限决定了特定角色可以使用哪些命令，或者更加准确的是，命令行接口（CLI）命令剖析树中的哪些节点。因此，用户可以创建一个名为“ no_debug ”的角色，它让分配到该角色的用户可以执行除了所有调试命令以外的所有命令。准入系统的精确度可以达到剖析树的第二级。因此一个角色甚至可以被定义为“ no_debug_fspf ”，它让用户可以执行除了 FSPF 调试命令以外的任何系统命令——包括调试命令。用户可以利用 CLI 命令在本地交换机中定义和分配角色。角色的分配甚至可以集中于某个 Radius 服务器，以便于管理。系统提供了两个缺省的角色，即网络管理员（完全访问权限）和网络操作员（只读访问权限）。用户最多可以定义 64 种角色。只有网络管理员级别的用户才能创建新的角色。

图 8　Cisco MDS 9500 系列的基于角色的访问



总结

存储网络的停机可能会对整个企业的基础设施造成严重的影响。这可能会使企业每年损失数百万美元。通过设计一个强大的、高度灵活的存储局域网，用户可以大幅度减少甚至消除停机时间。Cisco MDS 9500 系列多层控制器可以提供硬件冗余和可靠性，获得 99.999%的硬件正常运行时间。除了硬件冗余性以外，Cisco MDS 9500 系列还可以提供具有很高弹性的软件和一组创新的高可用性功能，以消除存储网络的停机时间。