第9章 存储网络技术

SFP模块光信号强度知识介绍

原创 EMC中文技术社区 戴尔易安信技术支持 *2016-07-27*

FC（fibre channel）交换机使用光信号传输数据，交换机的SFP/GBIC模块负责接受/发送光信号，并完成光/信号的相互转换。如果SFP模块接受/发送的光信号强度不够，势必会影响到上层FC链路的稳定性。一个优秀的SFP/GBIC模块是FC链路稳定的最基本保障。本文为大家衡量光信号强度的方法，EMC推荐的正常光信号强度范围，以及如何在Cisco/Brocade SAN交换查看SFP光信号强度。

衡量方法:

业界常见衡量光信号强度方法有两种：Microwatts(mW)和dBm，不同平台交换机采用方式可能会不一样，部分会选择mW，部分会选择dBm。SFP光模块信号强度通常包含两个指标，分别是Tx Power和Rx Power。Tx Power代表SFP模块发送方向的光信号强度；Rx Power代表SFP模块接受方向的光信号强度。

mW和dBm之间联系

mW通过功率方式描述光信号强度，dBm是decibel of the measured power to one millwats的简称，通过分贝方式描述光信号功率比。Cisco交换机使用dBm方式，Brocade交换机使用mW方式。两者可以通过以下公式互相转换：

dBm -> mW：

$$mW = 10^{dBm/10}$$

mW -> dBm：

$$\mathrm{dBm} = 10 * (\mathrm{Log\,mW})$$

EMC推荐光信号强度范围：

microwatt	milliwatt	dBm	Description
1.0	0.0010	-30.00	Loss of Signal
10.0	0.0100	-20.00	
25.1	0.0251	-16.00	2Gbps最小接受信号
31.6	0.0316	-15.00	4Gbps最小接受信号
50.0	0.0500	-13.01	
100.0	0.1000	-10.00	2Gbps最小发送信号
125.9	0.1259	-9.00	4Gbps最小发送信号
150.0	0.1500	-8.24	
200.0	0.2000	-6.99	信号强度可使用范围
250.0	0.2500	-6.02	
300.0	0.3000	-5.23	
350.0	0.3500	-4.26	
400.0	0.4000	-3.98	

常见速率最大可接受光衰减范围：

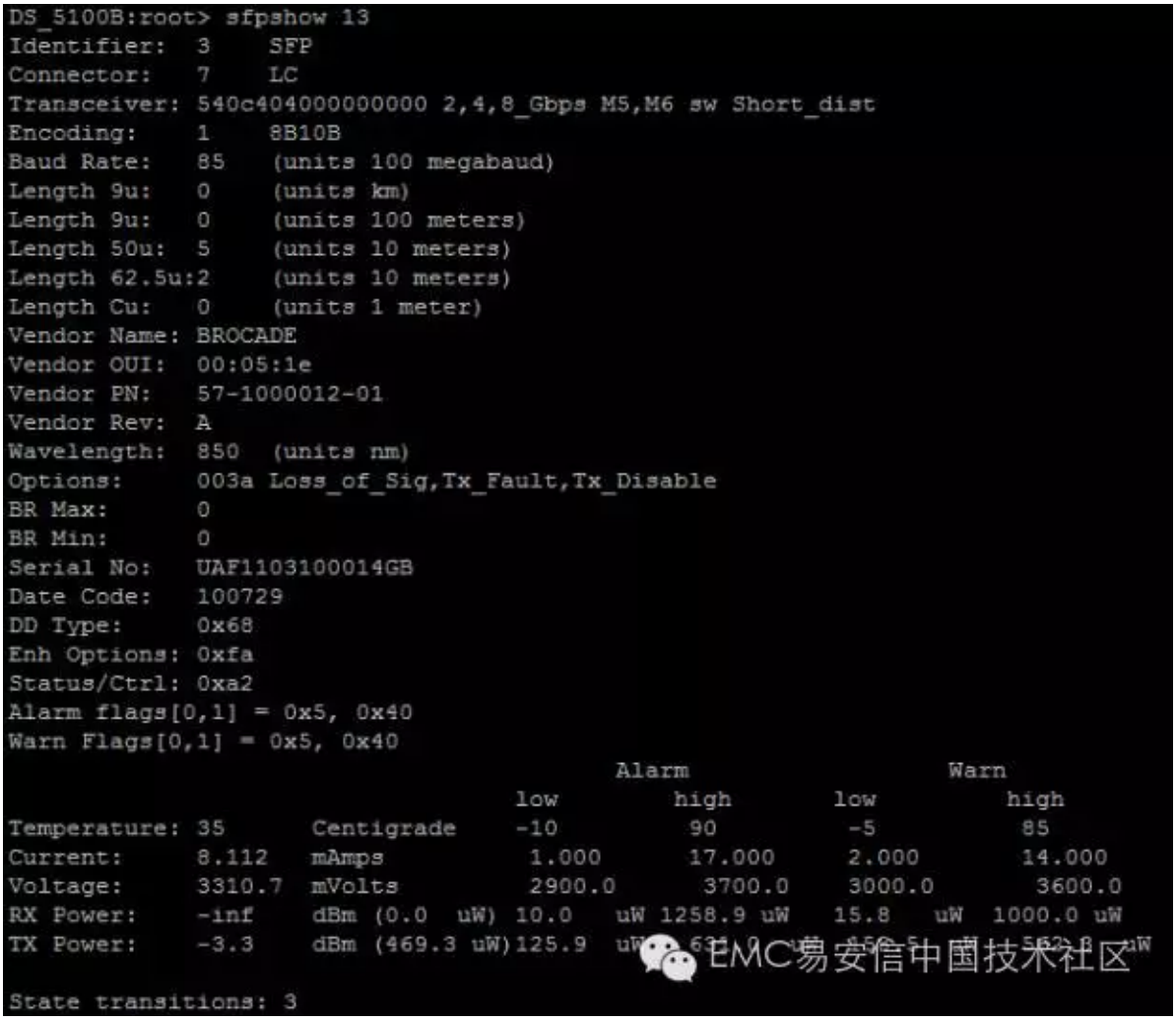
- 8Gbps最大可接受信号衰减值：-13.8dBm
- 4Gbps最大可接受信号衰减值：-15.4dBm
- 2Gbps最大可接受信号衰减值：-18.2dBm

查看SFP模块光信号强度：

Cisco/Brocade SAN交换机都提供工具用于查看SFP模块详细信息，包括SFP速率、序列号、Part Number、接受/发送方向光信号强度。

Cisco查看sfp模块光信号强度方法 – show interface transceiver details

Brocade查看sfp模块光信号强度方法 – sfpshow



从上面命令输出结果可以看出，Cisco/Brocade对光信号强度表示方法不一样，而且都提供当前信号强度，SFP有效光信号强度范围。只要当前SFP模块在有效范围以内，就可以认为SFP处于正常工作中。

EMC的推荐范围比Cisco/Brocade交换机自带范围要小，最小信号强度相对高一些，最强信号强度相对低一些。

SAN管理入门系列（一）交换机管理工具

原创 EMC中文技术社区 戴尔易安信技术支持 *2016-08-04*

简单介绍 EMC SAN交换机管理方法，以及常见管理工具。本系列共有8个章节。

**
**

EMC connectrix B系列交换机:

B系列交换机是指EMC贴牌生产的Brocade公司SAN交换机。主要有DS，MP和ED三大系列产品，分别对应中低端，多协议路由和高端交换机产品。

B系列交换机管理工具包含两大类型：

- **命令行**
    - 命令行可以通过IP网络或者串口线输入，也可以用来自动化管理大量的SAN交换机。
- **图形界面**
    - 图形界面有2种方式：一种是交换机自带的web服务器提供的web tool；另外一种是需要在单独安装的软件connectrix manager实现。
    - Webtool由交换机自带的web服务器提供，直接通过浏览器访问交换机管理IP即可。
    - Connectrix Manager软件是一个B系列交换机集中式管理工具，需要购买单独的license，可以管理多个不同型号B系列交换机产品，也提供些webtool不具备的功能。
    - Webtool和Connectrix manager都需要再本地安装Java程序。

**
**

备注:

图形工具并不能包含所有的命令行工具，很多操作只能通过命令行去完成。

EMC connectrix MDS系列交换机:

MDS系列交换机是EMC贴牌生产的Cisco公司SAN交换机。主要包含9100、9200、9500三大些列，分别对应中低端，多协议路由和高端交换机产品。

MDS系列交换机管理工具包含两大类型：

- **命令行**
    - 命令行可以通过IP网络或者串口线输入，可以用来自动化管理大量的SAN交换机。
- **图形界面**
    - 图形管理工具包含两种方式：一种是交换机自带web服务器提供的device manager；另外一种是需要单独安装的软件connectrix fabric manager。

- Device manager由交换机自带的web服务器提供，直接通过浏览器访问交换机管理IP即可。
- Connectrix Fabric Manager软件是一个MDS系列交换机集中式管理工具，需要购买单独的license，可以管理多个不同型号B系列交换机产品，也提供些device manger不具备的功能。
- Device manager 和Connectrix Fabric manager都需要再本地安装Java程序。

备注:

图形工具并不能包含所有的命令行工具，很多操作只能通过命令行去完成。

管理工具对比:

厂家	命令行	图形界面	
		交换机自带	单独安装
Brocade	1.典型的Linux模式，命令行可以自动补全。 2.Help输出交换机所有命令的描述，Help +"命令"查看一个命令的帮助。	webtool	Connectrix Manager
Cisco	1.典型的Cisco树状命令格式，命令行可以自动补全。 2.在命令行无法查看命令帮助，指定命令可以通过+" ? ".查看该命令参数描述。	device manager	Connectrix Fabric Manager
注意： 1.两个厂商的图形管理工具不能互相兼容。 2.交换机自带的图形管理工具无法完成VSAN，ZONE的配置操作。			

单独安装软件需要license，都支持SNMP 。

Link Aggregation(链路聚合)

原创 EMC中文技术社区 戴尔易安信技术支持 *2016-07-27*

在过去的10年里，以太网线速（Line Rate）的发展从10Mb/s -> 100Mb/s -> 1Gb/s -> 10Gb/s（40GE和100GE也已出现），而然有的时候single 10GE link依然无法满足VLAN Trunk或iSCSI/FCoE流量对带宽和冗余性的需求。于是链路聚合（Link Aggregation，LA）技术出现了，这是一种将多个网络链路合并成单条逻辑链路，从而提供更大带宽和冗余的网络技术。

注：本文多次用到Link Aggregation、Port Aggregation、Port Channel、Link Channel等术语，如未特别指出，可认为是同一个意思。

Link Aggregation

不同厂商交换机的端口聚合（Port Aggregation，PA）采用不同术语，Cisco的EtherChannel，Brocade的Brocade LAG，还有基于标准的IEEE 802.3ad LACP（Link Aggregation Control Protocol，该标准在2008年被转入IEEE 802.1ax），LACP可以动态配置端口聚合，且不依赖任何厂商，因此大部分以太网交换机都支持该协议。所有这些实现的目标都是一致的，即将两个或多个端口绑定在一起作为一个高带宽的逻辑端口来提升链路速度、冗余、弹性和负载均衡。技术上来讲，我们可以在交换机之间使用多个端口创建并行trunk链路，但生成树协议（STP）将其视为环路，从而会关闭所有可能造成环路的链接。端口聚合生成single logical link，不会造成环路，可作为Access Port（连接主机）或Trunk Port（承载Multi-VLAN traffic）使用。

在使用LA技术之前，有必要了解如下技术属性：

- 兼容性/正常运行：聚合链路每一端的交换机或主机必须理解或使用公共的端口聚合技术
- 负载均衡：通过哈希算法在single link中区分不同的traffic pattern来实现
- 链路冗余：如果逻辑链路中的一条物理链路故障，流量转走临近的链路，故障转移时间一般小于几个毫秒。一旦失败链路恢复，流量会重新分布。

在选择端口做端口聚合或端口隧道（port channel）时，每个端口需要互相兼容，可以在允许将端口加入端口聚合组（Port Channel Group）之前，检查端口的运行属性，兼容性检查通常包括以下接口运行属性：

- Port Mode
- Access VLAN
- Trunk native VLAN
- 允许的VLAN列表
- Speed

Link Aggregation Control Protocol (LACP)

LACP是通过向所有启用LACP协议的链路发送LACPDU Frame来工作的，如果发现链路另一端的设备也启用了LACP，LACP将独自在同一条链路上发送Frame，使得两者能够发现它们之间的多条链路，并将它们合并成单条逻辑链路。

LACP可以配置为两种模式中的一种：Active或Passive。在Active模式下，LACP主动在配置的链路上发送Frame；在Passive模式下，LACP的反应是“speak when spoken to”，从而可以作为控制意外环路的一种方法（只要其他设备在Active模式下）。

IEEE 802.3ad定义的LAC（Link Aggregation Group）是一个允许交换机自动协商端口绑定链路的协议，通过发送LACP Frame给Peer来实现。这些Frame在支持port channel的交换机端口之间进行交换，从而学习【邻居身份】以及【port group capability】并与本地交换机对比，然后LACP为port channel的端点分配角色。

系统优先级最低的交换机根据端口优先级决定哪些端口可在某一时段作为端口聚合的Active port。例如，一组8 link的port channel，在任何时候LACP选择4个优先级最高的端口作为Active port，通常数值越小的端口优先级越高。另外8条链路置于hot-standby状态，如果一条Active Link down，就会激活其它的链路。端口优先级是可配置的，如果没有配置，则使用不同厂商自己的默认值。如果端口使用了相同的值，厂商通常会实现一个“tie breaker”，比如lower port number作为Active port，即port 1/1 > port 1/5。

Port Channel中的每一个端口必须分配同样且唯一的channel group number，LACP自动在【配置为使用LACP的端口上】配置一个等于channel group number的管理键值（Administrative key value），该管理键值定义了端口与其他端口聚合的能力（ability）。一个端口与其他端口做聚合的能力取决于带宽、双工模式、点对点或共享介质的状态。通道协商必须设置为ON（无条件通道；没有LACP协商），Passive（被动监听并等待询问）或Active（主动询问）。

SAN网络设计原则

原创 EMC中文技术社区 戴尔易安信技术支持 *2016-07-23*

最近在研究Brocade相关硬件产品，在读到其中一篇SAN架构相关的白皮书时，里面有写到数据中心SAN网络的一些基本设计原则，在这里与大家分享下：

数据中心SAN设计大部分常见参数包括：

- 可用性 — 存储数据必须始终可被应用所访问到
- 性能 — 可接受的、可预测的、一致的I/O响应时间
- 效率 — 不浪费任何资源(端口、带宽、存储、电源)
- 灵活性 — 优化数据路径以有效利用容量

- 可扩展性 — 随时按需增加连接和容量
- 可服务性 — 加快故障排除和问题解决
- 可靠性 — 在SAN中设计的冗余且可靠的操作
- 可管理性 — 优化传输和存储管理
- 成本 — 设计费用控制在预算内，掌握实时运营支出

实际上，这些基本参数的适应范围可能依据客户的不同、职能SAN部署的不同而有所不同。一款经深思熟虑的SAN设计可综合考虑到所有这些因素，遵循博科SAN设计原则将有助于您协调不同需求之间的矛盾。此外，即便是复杂的大型数据中心SAN也可从一个崭新角度中获得收益。只有从这些基本需求着手来分析现有基础设施，这样才能找出其中能采用新SAN设计加以解决的差距及弱点，而在分析的同时仍可重新规划现有的基础设施组件。

原则#1: 最小化所管理Fabric架构的数量

原则#2: 最小化每个Fabric架构中交换机数

原则#3: 限制Fabric架构规模 (控制节点连接数量约在1,000到2,000个之间)

原则#4: 使用RAS(可服务性)水平高的交换机

原则#5: 避免过载比，以免造成拥塞或性能降低 (过载比比率在通过Fabric架构的所有相关数据路径时都应是一致的)

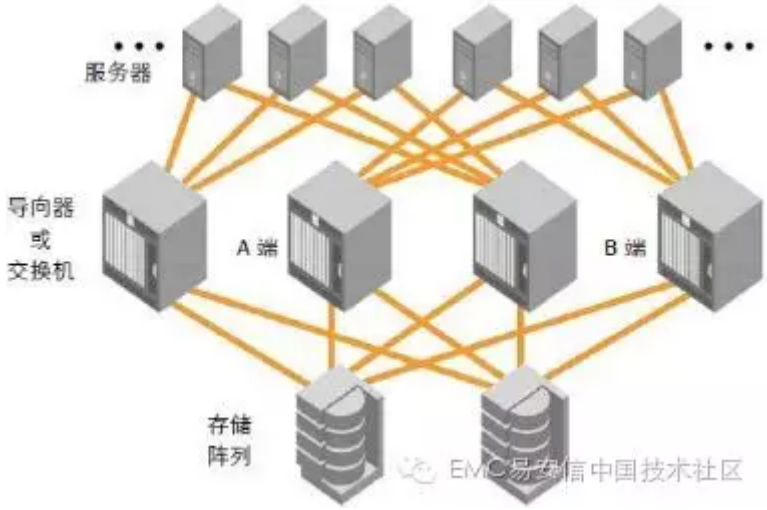
原则#6: 大型环境采用核心-边缘模式 (见附录)

原则#7: 针对存储流量进行设计

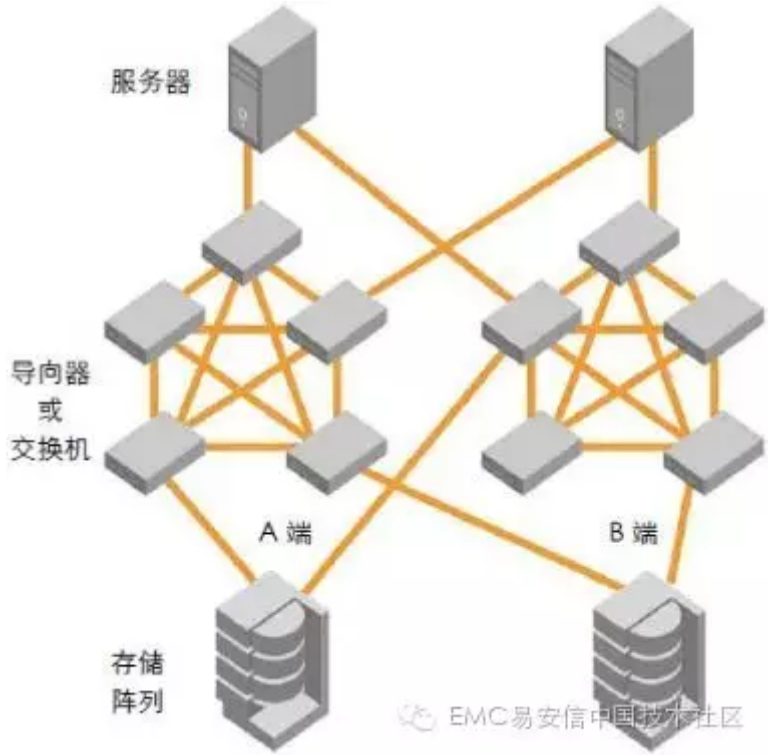
原则#8: 保持简单

附：三种SAN拓扑结构：

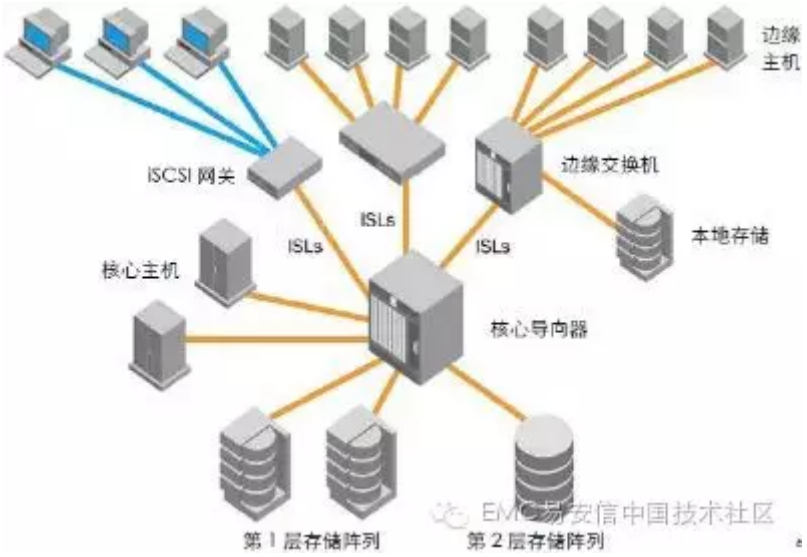
\1. 扁平SAN拓扑结构：



\2. 网状SAN拓扑结构：



\3. 核心-边缘(Core-Edge) SAN拓扑结构：



想了解更多信息的话可以去Brocade官网下载白皮书。

存储区域网络(SAN)中各种缓存(Cache)技术的应用和比较

原创 罗宾 戴尔易安信技术支持 *2016-07-28*

本文主要讨论了在SAN存储领域里Cache技术的应用范围和功能比较，可以为企业存储管理员等IT技术人员在制定企业级存储解决方案时提供参考。文章主要以EMC的CLARiiON 和 Symmetrix 等主流SAN存储产品为例，简要介绍SAN里涉及到的各种Cache缓存技术，同时对比各个技术的优缺点和应用场合，包括缓存使用中的最佳实践分析。由浅入深地阐述了缓存技术在后端存储阵列以及前端服务器层面越来越广泛的应用和多样化的设计与实现。最后引入了在实际生产环境下的使用缓存技术来提供生产效率的案例。

（1）引言：

在数据存储技术日新月异的今天，信息量也在以爆炸形式而迅猛增长。当磁盘等存储媒体介质已经不是存储行业的软肋时，大数据（Big Data）和云存储（Cloud Storage）便成为可能。各大存储厂商和研究组织也逐渐把新的焦点集中在如何提高数据存储速度和性能，从而进一步保证各种应用程序的读写达到更快速更可靠的要求得以满足。于是在当下流行的存储区域网络(SAN) 领域 ，新的提升存储性能技术就应运而生了。其中Cache缓存技术的发展和应用显得越来越重要了。

（2）正文：

在过去的十年中，服务器处理技术继续发展沿着摩尔定律曲线。每18个月，内存和处理能力已经翻了一倍，但是磁盘驱动器技术还没有。驱动器的转速，继续以相同的速度旋转。这导致了一个在I / O堆栈上的瓶颈，服务器和应用程序有能力以处理比磁盘驱动器可以交付的更多的I / O。这被称为I / O差距。传统的比较直接的方法是增加存储端处理器的DRAM主内存，从而能分配给控制器的读缓存（Read Cache）和写缓存（Write Cache）更多的容量，来加快后端存储的I/O吞吐量，这样就把处理I/O的任务交给了Cache和磁盘之间的通信。这里以EMC 的CLARiiON为例，首先介绍一下磁阵系统的读写缓存的概念。

一个存储系统的控制器缓存有两个部分：读缓存和写高速缓存。读缓存使用一个预读机制， 允许存储系统从硬盘预取数据。因此，应用程序时需要此数据就可以从读缓存里直接得到。存储系统内存储最近访问的数据的高速内存。为了得到更快的响应时间，所有主机写入都直接写入缓存，并在写入磁盘前向主机发回确认。写高速缓存的缓冲区，通过吸收高峰负荷优化写操作，联合小写操作，以及消除重写。你可以改变读缓存大小、写缓存大小和缓存页面大小达到最优性能。最好的尺寸的读和写缓存取决于读/写比率。一个通用的标准比读书写的都是两读/一写；即：读取占了66%的I / O。

在主存储DRAM(内存)的部署方式中，由用户决定在固态磁盘中放置哪些数据以及何时放置。用户必须进行特殊的操作来将数据迁移到固态磁盘中，而且使用到这些数据的应用程序必须被告知数据的准确位置。内存和缓存的固态 磁盘部署有两个显著的不同：在内存的部署方式中，仅有数据放置在固态磁盘上的应用程序可以提升性能。而且和缓存系统中性能随时间线性增长不同，内存部署方式可以即刻改善性能，不需要花时间预热数据。主内存分配给读、写缓存的空间越大系统处理I/O的速度就越快。这就是主内存DRAM的部署的最大优势，可以直接反应出对于来自服务器的I/O处理能力。提升对所有负载整体性能，见效快但是成本高，灵活性和可扩展性以及易维护性交差一点。

读、写缓存固然对提高存储系统整体性能很好，但是这些缓存依然来自存储控制器里的主内存，也就是DRAM。而由于内存硬件容量的制约和成本的考虑，基于RAM的系统主内存不可能无限增加，并且更重要的一个因素是，如果在使用大写缓存的情况下出现断电、宕机等不确定因素容易造成写缓存无法访问甚至数据丢失。那么就会出现缓存越大丢失的数据越多的情况。在实际生产环境中，很多客户都是因为突然的掉电导致系统内存中有来不及写入磁盘的数据而出现Dirty Cache（脏数据）如果后端存储的两个控制器都出现问题无法启动，那么这部分的缓存数据就彻底丢失了。还有就是很多核心硬件故障也会导致写缓存首先自动被禁用（Disabled），以便保证不会出现数据在写缓存内丢失的情况。

所以我们必须根据生产应用环境和以及I/O量大小及I/O特性来调整读写缓存的大小，来达到优化性能的目的。由此看来，单靠增加系统控制器的读写缓存，即单一增加系统主内存DRAM来提高存储系统整体性能还是不足的。因为内存的固态磁盘部署方式最大的缺点在于，今天最合适的数据分布方式可能并不是明天最合适的。举例来讲，假设一个非常关键的应用仅在每个月 末需要高性能，其数据必须在每月月末处理开始之前迁移到固态磁盘，并在月末处理结束之后移出。为解决这一问题，许多固态存储技术的供应商为其内存部署方式自动化软件提供自动化功能，可以自动化辅助选择和迁移数据到固态磁盘上。这些解决方案可以工作在全LUN级别或次LUN级别。此外，这些解决方案通常提供了基于策略的数据迁移功能，用户可以设置相关阈值，限制数据升级到固态磁盘上的次数和降级到普通磁盘上的次数。这些自动化分层软件解决方案目前都已应用成熟。

EMC的全自动分层存储（FAST）就是自动化分层软件解决方案中比较先进的技术之一。其实在SAN应用存储产品中，各类控制器，不论软件控制器，服务器内部的RAID控制器或者是高端外部阵列的控制器，会将固态存储技术作为前端传统磁盘存 储的一个缓存。该缓存控制器会区分出所有经常存取的数据，亦称为“热点数据”，并自动地将其迁移至固态媒介。虽然不同的缓存控制器或许有些许不同的缓存交 换算法，但最基本的想法还是通过将热点数据迁移到最高速的媒介上，提升I/O性能并降低I/O延迟来改善性能。I/O模式每时每刻都在变化，缓存控制器自 动地监控哪些数据是最常被访问的，并将其迁移至最高速的媒介，在这个过程中无需任何用户或管理员的介入。

另外FAST还具备Cache功能，如果存储系统有需求，简单讲就是Fast Cache将把一部分固态硬盘当成Cache来使用。FAST与 FAST Cache协同工作，从而基于应用程序访问各种数据块的方式将数据放置在最适当 的存储层上。自 EMC 在企业阵列的磁盘模块（通常称为 SSD）中首次部署闪存技术以来，拓展此技术在整个存储系统中的使用一直是业界的目标。IT技术日新月异，近几年里闪存技术的高性能和快速降低的每 GB 成本结合在一起导致缓存层 概念的出现。缓存层是使用企业闪存驱动器的大容量辅助缓存，介于存储处理器基于 DRAM 的主缓存和硬盘驱动器之间。在 EMC CLARiiON、Celerra 统一存储和 VNX™ 存储系统上，此功能称为 EMC FAST Cache。

FAST Cache 可扩展存储系统的现有缓存容量，以提高整体系统性能。具体通过以下几个方式实现：将经常访问的数据映射到闪存驱动器以扩展 DRAM 缓存的功能，闪存驱动器的速度比机械式驱动器（也称为硬盘驱动器）快一个数量级，从而大大提高系统性能。它还使用闪存驱动器提供更大的可扩展缓存，与 DRAM 容量相比，闪存驱动器可以提供非常大的每驱动器容量。FAST Cache 的容量介于 73 GB 到 2 TB 之间，这比现有存储系统的可用 DRAM 缓存大得多。无需用户干预即可使应用程序体验到 FAST Cache 的性能优势。因此，EMC用闪盘作为系统主内存Cache的延伸，不存在掉电后的保护问题。

使用 FAST Cache 的一个主要应用就是能提高应用程序性能，特别是对于 I/O 活动经常不可预知的大幅度增加的工作负载。应用程序工作数据集经常被访问的部分会复制到 FAST Cache，因此应用程序可以立即大幅提高性能。利用 FAST Cache，应用程序可以通过以闪存驱动器速度处理繁重的读/写负载来实现一致的性能。另一个重要好处是降低了系统的总体拥有成本 (TCO)，这是通过减少后端硬盘驱动器上的负载实现的。FAST Cache 将大型存储容量的繁忙部分复制到闪存驱动器；因此，许多 LUN 的最繁忙区域只使用一小组闪存驱动器。这使得一大组速度较慢的驱动器可以实现通常由速度较快的驱动器提供的

性能。经过一段时间后，速度较快的光纤通道驱动器可以减少数量或替换为速度较慢的光纤通道或 SATA 驱动器，同时保持相同的应用程序性能。这样就提高了存储系统的财务效率和能源效率 。

这里我举个实际生产案例来说明FAST Cache如何提高了应用程序的性能。原来在客户环境里有若干台虚机，针对150个VMware试图桌面，所有虚机启动需要20分钟，响应225毫秒，使用FAST Cache之后启动时间缩短为9分钟，响应时间缩短为50毫秒。还有我们知道Exchange Server 2010 是一个对存储系统要求很高的应用程序。而使用Exchange 2010的公司就会从EMC FAST Cache 中受益匪浅。我们做了测试，在一个单一Exchange 2010 服务器使用SATA盘作为储存的环境，启用FAST Cache测试发现IOPS获得了113%的增长。在另一个更大一些的配置达到44TB数据量的Exchange 2010环境，数据处理任务非常繁重，测试结果显示在启用FAST Cache后STAT 盘的性能也有42%的提升。

FAST Cache 与存储系统缓存的比较，FAST Cache 是基于半导体的存储技术。它在存储系统的 DRAM 缓存（速度较快但容量有限）与机械式硬盘驱动器（速度较慢但容量较高）之间提供一个基于闪存的大容量辅助缓存层。

DRAM 内存与 FAST Cache 缓存的比较

特征　DRAM 缓存　FAST Cache 缓存

位置　它最接近 CPU，并且延迟最低。　与 DRAM 缓存相比，它距离 CPU 较远并且较慢。

粒度　它具有非常高的粒度，粒度实际上是 I/O 大小。缓存页面大小可由用户配置，可介于2KB 到16KB 之间。　它在 64 KB 粒度的范围内运行。

可升级性　不可升级。　可在各种型号中升级，相关选项取决于存储系统型号和闪存驱动器类型。

操作　它对读操作和写操作分别使用单独的区域。　它使用单个区域来完成读操作和写操作。

容量　与 FAST Cache 相比，它的大小有限。　可以扩展到非常大的容量。

范围　它支持 FAST Cache LUN 以及其他 LUN。　它允许您在所选 RAID 组 LUN 或存储池上启用 FAST Cache。

响应时间　响应时间大约为几毫微秒到几微秒。　响应时间大约为几微秒到几毫秒。

可用性　出现故障时，需要合格人员进行更换。　出现故障时，另一闪存驱动器热备盘自动取代出现故障的驱动器，客户可更换故障组件。

电源故障　它的内容具有易失性；因此不能经受断电。　它的内容是非易失性的，可以经受断电。

如何提高提升应用程序性能是储存行业一个永恒的话题。最近几年闪存技术的兴起和飞速发展让很多业内人士又看到了新的机会。我们知道SAS和SATA驱动器给数据库提供了不凡的性能容量比，但基于物理旋转的磁盘注定不能够提供最佳的性能。 于是把闪存驱动器加入到磁阵中就发现可以提供一个更高数量级的性能。 现在在市场上有一个新的服务器闪存缓存技术，提供了更大的性能。就是说 把Flash闪存放置到装有PCIe卡的服务器上，就可以加速甚至得到另一个数量级闪存驱动器的性能。

EMC VFCache 就是一种优秀的服务器闪存缓存解决方案，可利用智能缓存软件和 PCIe 闪存技术缩短延迟并提高吞吐量，从而大幅提升应用程序性能。VFCache 可加快数据块 I/O 读取速度和保护数据，方法是使用到网络存储的直写缓存，提供持久的高可用性、可靠性、数据完整性和灾难恢复。与基于阵列的 EMC 全自动存储分层 (FAST) 软件配合使用时，VFCache 可创建从应用程序到数据存储区的最高效和最智能的 I/O 路径。并最终衍生出针对物理和虚拟环境的性能、智能和保护进行动态优化的网络基础架

构。VFCache 与阵列上的 FAST VP（Virtual Provision） 和 FAST Cache 互为补充，但它并不要求存储阵列使用 FAST VP 或 FAST Cache。VFCache 作为一种服务器端只读缓存，可将 EMC 全自动存储层 (FAST) 策略扩展到服务器。VFCache 动态缓存引擎不仅会自动识别频繁访问的数据，还会自动将该数据升级到 PCIe 卡上的闪存中。如此一来，不但使来自 SAN 和共享阵列的 I/O 压力得到缓解，而且也提高了应用程序性能。由于频繁访问的数据位于服务器内，使 I/O 服务次数大幅减少。因此，服务器可支持更多的虚拟机和/或应用程序可交付更多的事务。

VFCache 独立于 FAST Cache 和 FAST VP 运行。VFCache 是一种专用的服务器端只读缓存，而 FAST Cache 是一种支持读写操作的共享阵列端缓存。上述二者均属暂时性缓存，而 FAST VP 则是在基于阵列的存储池内的层之间半永久性地移动数据。借助于 EMC 存储阵列缓存和分层技术中的智能优势，FAST VP 算法允许阵列集中资源来处理所有工作负载中要求最高的部分。这三种技术旨在共同确保以最低的延迟为最常访问的信息提供服务。VFCache、FAST Cache 和 FAST三者相结合，可进一步简化数据连续体系间的数据传递。

同样，我们也找了个实际案例来说明VFCache如何提高了应用程序的性能。在一个使用VFCache 配合 FAST Cache 来加速联机事务处理系统（OLTP）Oracle DB11gR2的环境里，我们发现在配置了VFCache后，系统的吞吐量，每分钟交易量，是基线配置(Baseline Configuration)的2.5倍，而延时时间（Latency）却减少了30%。如果再同时配合启用FAST Cache功能，系统的吞吐量及每分钟交易量，是基线配置的7.8倍，而延时时间减少了20%。

（3）结论：

回头看来，总结我遇到的这些实际案例中，一些用户选择了固态磁盘缓存的部署方式，而另一些则选择主存储(DRAM)方式。由于每种方式都有其优点，这使得许多客户开始时选择缓存的方式，而之后又为其解决方案额外增加了内存方式的支持选项，反之亦然。主内存是基本配置，用户可以根据EMC官方文档找到自己型号机器所对应的最佳推荐读/写缓存配置。但是内存是不保护数据的，一旦断电数据将全部丢失。这将是它永远的软肋。当存储系统的FLARE® 版本 升级到30以上后，并且根据条件能够使用Flash 闪存驱动器，那么我们极力推荐用户使用EMC FAST VP及FAST Cache来进一步提高整体系统性，因为闪存驱动器的速度比机械式驱动器快一个数量级。FAST Cache 的容量介于 73 GB 到 2 TB 之间，这比现有存储系统的可用 DRAM 缓存大得多。但是FAST CACHE并非适用于所有I/O类型。例如，连续大I/O数据流或许根本不会促使数据被提升（Promote）至FAST Cache，因为这些I/O不会多次访问同一个64KB数据块（chunk）。于是人们开始把目光从后端的存储转移到了前端服务器。VFCache就是一个软硬件完美结合的最新服务器端闪存缓存解决方案。在内存中的应用缓存数据只能加速某个应用，但VFCache可以加速 所有连接到它的源存储，也就是我们通常所说的LUN。当然现阶段VFCache还不支持Cluster，希望能在下一版VFCache中实现了共享磁盘环境和Active/Active集群的支持。综上我们不难看出，主内存DRAM，FAST Cache 还有VFCache 都是提升存储性能的好办法，三者如果能够根据用户实际应用程序的特点以及硬件设备条件来搭配使用必将给用户带来前所未有的性能提升体验，同时也会降低运维成本，提高客户满意度和企业生产效率。

存储网络 – 了解FCoE的8个技术细节

原创 EMC中文技术社区 戴尔易安信技术支持 *2016-08-01*

Fibre Channel over Ethernet（FCoE）是一个把Fibre Channel（FC）中的帧（Frame）封装到一个增强的以太网（Ethernet）中的传输协议标准，它应用与组件存储网络。关于FCOE的介绍网络上非常多，但大多都比较分散。本文通过总结了8个关于FCoE的技术细节，将存储网络FCoE中必须要了解的知识进行整理。

1.FCoE就是用以太网来传输FC：

FCoE的全称是Fiber Channel over Ethernet，旨在通过以太网直接传输光纤协议，让存储网络中的数据可轻易跨越光纤和以太网的界限，通过同一种截物理介质（以太网络）进行传输，同时保留了FC中的上层协议的特性，例如数据一致性，流控制等，但不使用原先光纤网络的物理线路与接口。目的在于降低用户在存储网络构建和管理方面的成本和复杂性。

2.FCoE的优势是融合网络和未来高速带宽的预期：

**
**

FCoE的优势显而易见，部署FCoE以后，企业只需要使用以太网络构建数据中心的网络，而不是原来使用光纤网络和以太网络进行结合。同时减低线路的总数，主机端接口卡（HBA，网卡）等的总数。而这两者会被集成为同一个接口设备 –融合网卡Converged Network Adapter（CNA）用来同时处理FC协议和TCP/IP协议，从而保证在同一个主机接口上对存储网络和IP网络进行隔离。另外，未来FCoE可能会提供相比FC网络更大的带宽，FCoE目前起始就是用10 Gb的以太网，而40 Gbps和100 Gbps的以太网也相信在不久的将来也会推出，相对FC网络方面，8Gb和16Gb是主流，最新的32Gb也刚刚出现，但相对也在起步阶段的100Gb以太网来说还是稍逊不少。

3.FCoE主要协议还是FC，只是封装以后用以太网介质传输：

**
**

FCoE协议的发明目的很简单，用来把FC协议封装到以太网之中。下面一张图可以很清楚的清楚的看到一个FCoE的帧包含哪些部分。FC帧和以太网帧，在FCoE中的传输是一1：1的对应关系存在，没有任何封包和合并多个FC帧到一个以太网帧的情况。因此，在FCoE数据传输的每个节点上（网卡，交换机）都需要启用巨帧Jumbo Frame以支持封包以后的以太网帧的传输（以太网原来默认帧最大1500字节，传输FC帧需要2112）。不同的厂商的产品默认的巨帧MTU有所不同，比如思科的Nexus默认值是2158。

Ethernet frame with Fibre Channel frame insertion.

FCoE stack.

4.FCoE需要配合特定设备，且运行在增强的以太网上：

**

**

FCoE其实不能运行于普通的10/100Mb， 1Gb和10Gb的以太网络，因为普通的以太网并不是Lossless Ethernet，它不包含任何重传（重传由上层的TCP协议控制）和流控制技术， 保证传输等功能。FCoE所运行的以太网络是一种增强的以太网络（Enhanced Ethernet），可以包含这些功能。不同的厂商对这种增强命名有所不同，思科把它叫做Data Center Bridging（DCB），博科和其他厂商则把它叫做Converged Enhanced Ethernet（CEE）。

现有的支持1Gb以太网络的RJ45的CAT-5和CAT-6网线也不能支持FCoE的产品，转移到FCoE的用户可以选择现有光纤线和一种新类型的扁平线缆（Twin Axial）作为传输介质。FCoE同时还需要配合Converged Network Switch（CNS）作为连接以太网和光纤网络的桥梁

5.Enhanced/Lossless Ethernet比普通以太网增加了一些流控制协议；

**

**

原先的以太网使用到了一种叫做PAUSE的机制，PAUSE机制可以防止瞬时过载导致缓冲区溢出时不必要的帧丢失，实现了一种简单的停-等式流量机制，来提高传输的质量。而原有的PAUSE机制在决定对特定端口进行停-起的操作的时候并没有一个优先级的控制。增强的以太网解决这个问题，在原有的以太网的基础上增加了一些扩展的协议机制，使得让以太网更适合存储网络。这种控制机制主要包括以下几种：

- Priority Flow Control（IEEE 802.1Qbb）和Enhanced Transmission Selection（IEEE 802.1Qaz），它们的作用是可以对帧的传输优先级进行调整。例如将存储网络的帧的优先级调高，获得更高的带宽，以保证存储网络传输的速度和质量。
- Congestion Notification（IEEE802.1Qau）作用是对以太网中造成冲突的源和目标端同时叫停的机制。

另外，增强的以太网还提供了二层网络的多路径机制，替代原有Spanning Tree Protocol（STP）只支持单路径。

6.FCoE网络中的端口和FC网络类似，只是叫法不同：

- N_Ports（HBA和存储FC前端口）在FCoE中叫做VN_Ports（CNAA和存储FCoE端口）
- F_Ports（Fabric端口）在FCoE中叫做VF_Ports（FCoE交换机的端口）
- E_Ports（FC交换机到FC交换机）在FCoE中叫做VE_Ports（FCoE以太网交换机互联的端口）

7.FCoE应用场景是对现有存储架构的扩展，而不是作为iSCSI或者NAS的替代：

**
**

如果你的企业在以太网上部署存储，如果你的规模不是很大，iSCSI和NAS都是在BLOCK和FILE端很好的选择。那如果您的企业已经有更大规模的存储网络，在新建基础架构的情况，使用融合的FCoE网络可以同时兼顾BLOCK和FILE存储，10Gb的速度可以同时用来支持FCoE，iSCSI，NAS，将来以太网的高速发展也会使得企业在部署FCoE上得益。

8.EMC的存储产品全面支持FCoE：

EMC的存储产品已经从2010开始逐步支持FCoE网络，EMC支持FCoE的的产品列表参考：EMC产品的FCoE的兼容性列表

存储网络 – FCoE与FCIP间的小同大异

原创 EMC中文技术社区 戴尔易安信技术支持 *2016-08-02*

FCoE和FCIP都是将Fibre Channel（FC）协议运行于非光纤网络载体的存储传输协议。它们将FC协议封装在网络OSI模型的不同层中，这种封装技术上实现的区别使得FCoE和FCIP在应用和部署上有所不同。本文就FCoE和FCIP区别入手，描述FCoE和FCIP两种存储网络协议的共同点和，区别，以及对应的不同应用场景。作为读者在选择网络存储架构的参考。

FCoE的全称是Fiber Channel over Ethernet，是通过以太网直接传输光纤协议，让存储网络中的数据可轻易跨越光纤和以太网的界限，通过同一种截物理介质（以太网络）进行传输，同时保留了FC中的上层协议的特性，例如数据一致性，流控制等，但不使用原先光纤网络的物理线路与接口。目的在于降低用户在存储网络构建和管理方面的成本和复杂性。（参考：存储网络 – 了解FCoE的8个技术细节）

FCIP的全称是Fibre Channel over Internet Protocol，是通过运行于以太网中的IP通道来传输FC协议，使得可以将两个距离比较远的FC SAN网络连接在一起，组成一个更大的存储网络。它的好处是无需在两地之间再部署光纤网络，就可以实现使用现有的IP网络传输FC协议。

FCoE和FCIP在应用上有一些共同点：

- 都可以利用现有的以太网/IP网络对存储网络进行扩展与融合。
- 都是把FC协议封装在不同的数据包中进行传输，而FC协议中内部都是SCSI命令。
- 都需要额外购置专用的设备，FCoE需要添置FC和以太网融合的交换机支持无损以太网传输FC协议，主机端需要配置融合网卡。FCIP则需要在两个SAN网络之间IP通道IP网管支持拆解包的操作（例如Cisco MDS的IP Storage模块）。
- 都可以利用现有的高速以太网满足带宽上的需求。

FCoE和FCIP在应用上的主要区别有以下几点：

- 传输方式方面，FCIP是将FC包交由OSI第四层的TCP/IP包封装，FCoE则是直接封装在二层的以太网帧内。FCIP直接使用第三层实现处理拥堵，服务质量和传输优先权。FCOE则依赖于增强的无损以太网实现流控制，传输优先权等等。
- 由于FCIP和FCoE在封装的方式不同，FCIP可以直接使用TCP/IP网络中的路由功能，FCoE则只能通过MAC来寻址，同时管理员需要进行点对点基于WWN的Zoning配置。
- FCIP仅仅应用交换机到交换机的连接，而FCoE则是主机 – 交换机 – 存储的多方位连接。
- FCIP对以太网没有速度上的限制，而FCoE则需要起始10Gb的带宽，且是无损以太网。
- FCIP通常来说适合长距离的SAN网络的互联，数据传输，容灾需求，IPSec功能还可以为FCIP的传输进行加密。而FCoE则更适合数据中心能的网络融合，满足数据访问需求。

总结来说，FCoE和FCIP技术的出现，都是为了在已有的以太网和IP网络中使用FC来传输SCSI，让以太网可以用作存储网络之用。无论是使用FCoE基于数据中心内部和融合网络，以及FCIP实现数据中心间的存储网融合。主要的优势还是体现在成本上，虽然在维护和管理这些网络与设备也需要额外的工具与人力，而相对于节省下来维护两个独立的光纤网络和以太网所产生的成本，或者说是就未来的存储网络规划来说都是有一定的优势存在。

SAN网络迁移指南——Brocade篇

原创 EMC中文技术社区 戴尔易安信技术支持 *2016-07-26*

如今的SAN网络管理员面对是不断增长的存储和吞吐量需求。这需要既可以满足当前需求，也可以应对未来增长的高性能和全冗余的SAN网络。为了适应这些变化，SAN管理员们不得不经常考虑迁移或升级现有的存储网络。

上一次我们介绍了SAN网络迁移前的准备工作，这回我们将介绍将SAN网络迁移至Brocade光纤交换机的步骤。

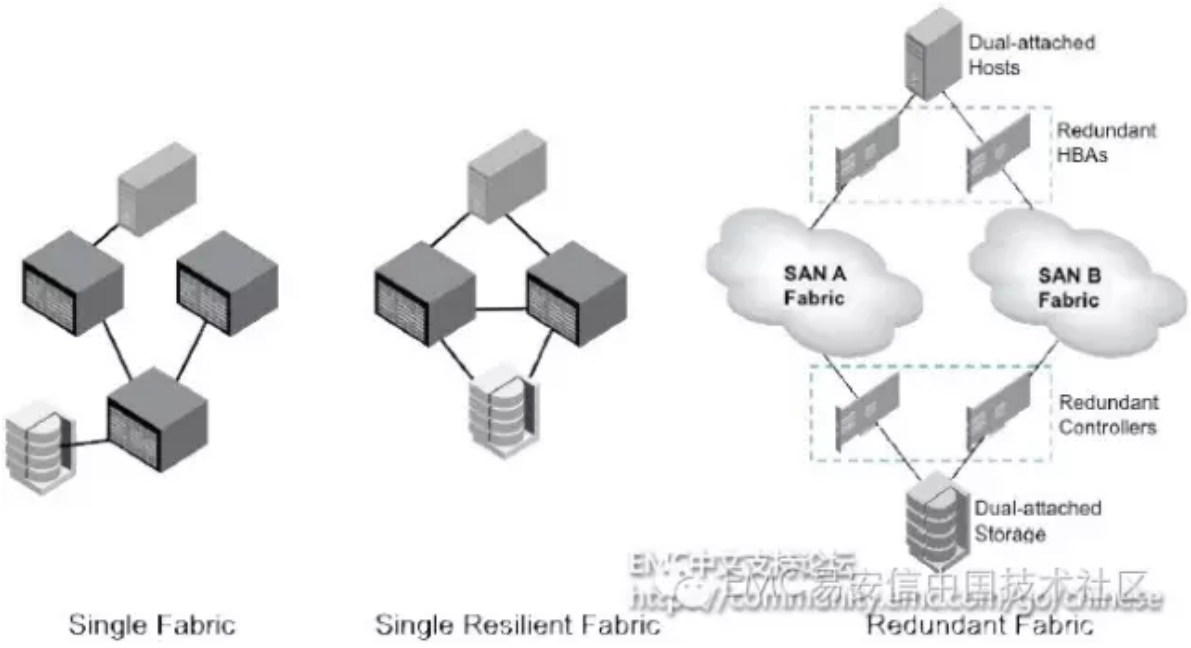
SAN网络迁移指南——迁移前准备工作

迁移前评估:

在正式迁移光纤网络前，了解当前应用环境和新网络需求是非常重要的。取决于现有SAN网络的架构、光纤网络拓扑、大小和活动设备的数量，具体的迁移过程有不止一种方法。取决于应用和项目的需要，SAN网络还可以在线或离线的方式来完成迁移。因此，不管是哪一种迁移方法，下面这些因素是需要考虑的：

- 评估当前的光纤网络拓扑
    - 应用故障转移需求
    - 存储故障转移需求
    - 是否需要同时升级现有拓扑（去除SAN网络中性能瓶颈、服务器/存储扩容、布线管理等常规光纤网络维护）
    - Zone配置导出方法
    - 服务器、存储物理摆放位置优化
- 评估新光纤网络拓扑
    - Brocade Fabric OS升级要求
    - 捕获现有交换机配置并与Brocade交换机配置比较
    - 导入Zone配置方法
    - 中继（Trucking）设置
    - 未来服务器、存储物理扩展规划
- 硬件安装逻辑规划
    - 机架空间需求

    - 电源需求
    - 布线需求
- 拓扑和Zone配置规划
- 初步的迁移计划



迁移策略：

在评估完所有需求后，迁移步骤主要分为下面两大类：

- 在线迁移（冗余的光纤网络）
    - 冗余的光纤网络使得在升级或迁移部分设备时，将活动I/O重定向到另一个光纤网络成为可能。整个迁移活动不影响当前的I/O操作。使用这种策略，主机在降级模式下工作且没有多余的数据路径提供故障保护。通过适当的规划，可以最小化停机或中断时间。一旦完成网络的升级并验证，通过复原数据路径，可以迅速重新上线
- 离线迁移
    - 光纤网络还可以离线迁移，在维护时间内，停止主机使得前端I/O完全停止。这是最安全且最方便的迁移方法，当然前提是应用允许停机。

迁移计划和准备：

一个良好的迁移计划包含下面这些步骤：

- 项目范围和成功条件
- 阶段、任务和子任务
- 资源定义

- 时间线安排
- 任务依赖性
- 跟踪和检查点
- 流程、设计和配置的移交
- 回滚计划和中止条件

具体到迁移，可以从下面几个方面来检查：

迁移步骤	结果	注释
机柜、机架、线缆和电力部署		
安装Brocade Network Advisor		
安装Brocade SAN Health并搜索Brocade和旧有光纤网络设备		使用SAN评估和Zone导入工具
安装配置Brocade交换机的管理口和串口		
安装推荐的Brocade FOS		
创建所有交换机的基准配置		
导入Zoning设置		使用Brocade SAN Health
安装授权许可文件		
删除不再使用的Zoneset		
为新设备创建Zone		Brocade HBA可以使用Dynamic Fabric Provisioning功能
验证ISL配置		Spinfab和D-port（16Gbps平台）

迁移和验证：

迁移完成后，很重要的一步是迁移后的验证。下面这些步骤可以确保哦哦所有的任务正确完成：

- 运行Brocade SAN Health工具
- 确认新的SAN网络配置
- 确认应用工作状态
- 完成SAN迁移收尾工作
- 撤除旧有设备

SAN网络迁移指南——迁移前准备工作

原创 EMC中文技术社区 戴尔易安信技术支持 *2016-07-25*

如今的SAN网络管理员面对是不断增长的存储和吞吐量需求。这需要既可以满足当前需求，也可以应对未来增长的高性能和全冗余的SAN网络。为了适应这些变化，SAN管理员们不得不经常考虑迁移或升级现有的存储网络。

在不同供应商的设备间进行迁移往往需要更多的工作。从设计、配置一直到部署流程和迁移后的审计等。本文将介绍将迁移现有SAN网络设备前的准备工作。关于迁移Cisco和Brocade交换机的具体步骤，将在后续的文章中介绍。

迁移方法和主要步骤:

总的来说，迁移一个SAN网络有下面三种方法：

重新构建（Rip and Replace）：顾名思义就是简单地直接替换新交换机

逐步替换（Cap and Grow）：新设备以并行的方式与现有SAN网络的设备同时工作。新主机或部分现有主机接入新交换机，旧交换机逐步地被替换，知道旧SAN网络中不再有设备、

互通（Interoperate）：新交换机是交互模式（Interop Mode）与旧有设备一同工作。

迁移的大致步骤如下：

准备：分析当前的存储架构、业务需求和分析。识别关键服务器、存储系统和应用。准备好回滚（Rollback）方案以防万一。准备或升级SAN网络架构图。提前准备好所有设备的配置（包括Zone、VSAN配置等）。根据不同的迁移方法，大多数的配置可以在实施迁移前决定好。

计划和分析：决定迁移方案并创建迁移策略。识别新增的或未来的SAN网络需求。这一步的目的是为了让迁移后的网络具备足够的灵活性。

实施和优化：实时迁移，移动网线和机柜，配置设备。在迁移完成后可以实施持续的监控和优化以减少风险和让SAN网络适应新的项目和应用需要。

迁移前准备：

任何迁移动作都需要准备工作，包括：

- 列举网络设备清单
- 确认兼容性
- 升级组件（以满足设备兼容性）
- 评估SAN网络
- 验证迁移后应用状态

规划和设计：

规划和设计环节主要包括物理和结构设计两部分：

**
**

物理规划

物理规划包括物理空间确认、空调系统、电力系统、PDU（Power Distribution Unit）电源排插、布线和理线面板等内容。不同的厂商对机柜有不同的需求，可以去厂商的支持站点下载设备硬件和环境要求。电力规划还需注意的是设备所需的直流电（DC）和交流电（AC）数量。

结构规划

结构规划包括所有设计有关的细节，包括网络拓扑、布线图、布线技术、电源排插位置、布线方式、空调和风量计算等。

软件交互性（兼容性）规划

在迁移SAN交换机时，交换性（Interoperability）是很重要的考量。当不同厂商的交换机在同一个Fabric网络中一齐工作时，软件交换性就扮演着重要的角色了。关于Brocade和Cisco MDS SAN交换性的交互性设置，请参考文章：

EMC光纤交换机互通性管理

**
**

软件授权

在迁移SAN交换机前，需要确认新交换机和所需功能的授权都已正确安装。通常大多数功能都已涵盖在基础授权（Base License）中，但一些特定的应用仍然需要分别购买并安装。可以去厂商的支持站点下载所需功能所对应的授权名称。