第1章 网络存储主要技术

NAS简介

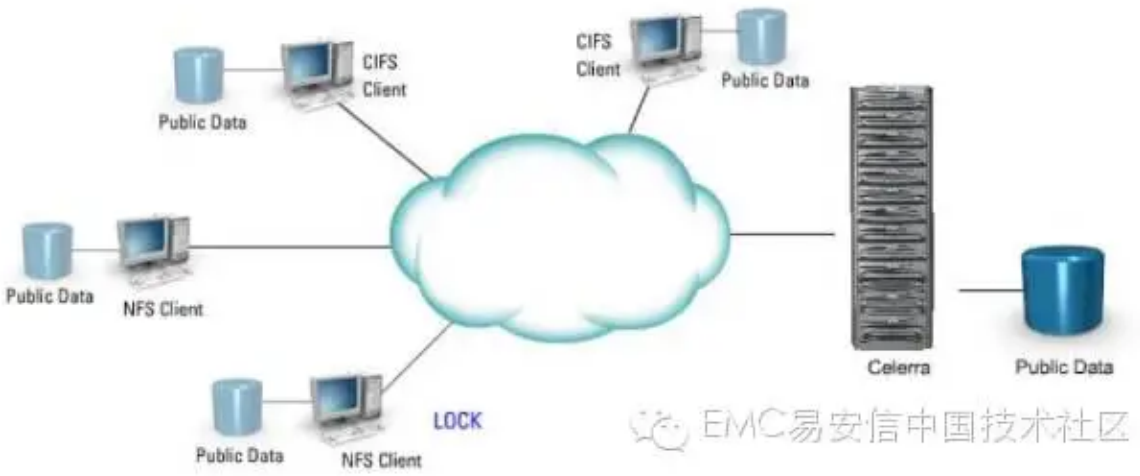
原创 EMC中文技术社区 戴尔易安信技术支持 *2016-03-07*

在20世纪80年代初，英国纽卡斯尔大学布赖恩.兰德尔教授 (Brian Randell)和同事通过"纽卡斯尔连接"成功示范和开发了在整套UNIX机器上的远程文件访问。继"纽卡斯尔连接"之后， 1984 年Sun公司发布了NFS协议，允许网络服务器与网络客户分享他们的存储空间。90年代初Auspex工程师创建了集成的NetApp文件管理器，它支持windows CIFS和UNIX NFS协议，并有卓越的可扩展性和易于部署，从此市场有了专用NAS设备。在短短几年中，NAS凭借简便高效应用的中心思想，逐渐成为网络数据存储方案的主打设备之一。目前EMC公司 Celerra产品拥有优异的性能及多功能性，在全球NAS市场处于领导地位。

NAS概念

NAS（Network-Attached Storage，网络附加存储）是指连接到计算机网络的文件级别计算机数据存储，可以为不同客户端提供数据存取。

NAS被定义为一种特殊的专用数据存储服务器，包括存储器件（一个或多个硬盘驱动器的网络设备，这些硬盘驱动器通常安排为逻辑的、冗余的存储容器或者RAID阵列）和内嵌系统软件，可提供跨平台文件共享功能。NAS通常在一个LAN上占有自己的节点，无需应用服务器的干预，允许用户在网络上存取数据，在这种配置中，NAS集中管理和处理网络上的所有数据，将负载从应用或企业服务器上卸载下来，有效降低总拥有成本，保护用户投资。



NAS本身能够支持多种协议（如NFS、CIFS、FTP、HTTP等），而且能够支持各种操作系统。NAS是真正即插即用的产品，并且物理位置灵活，可放置在工作组内，也可放在混合环境中，如混合了Unix/Windows局域网的环境中，而无需对网络环境进行任何的修改。NAS产品直接通过网络接口连接到网络上，只需简单地配置一下IP地址，就可以被网络上的用户所共享。

NAS特点

与采用存储区域网络(SAN-Storage Area Network)的方案比较，采用网络附加存储(NAS-Network-Attached Storage)结构的方案具有以下特点:

1. 以网络为中心，开放的标准协议支持

区别于存储区域网络(SAN)的设计方案，网络接入存储(NAS)的模式以网络为中心。该方案利用现有的以太网网络资源来接入专用的网络存储设备，而不是另外再部署昂贵的光纤交换机网络来连接传统的存储设备，这样保护了用户对以太网的投资。

近年来，千兆以太网的传输带宽(1000Mbps，为125MB/s)已经得到普及，并且有望朝万兆以太网发展。届时，以太网的传输带宽将会是10倍于SAN赖以生存的各种SCSI和 Fiber Channel协议的传输带宽。EMC公司Celerra产品支持目前最流行的TCP/IP网络协议，而使用的NFS和CIFS文件服务协议也是业界标准协议，充分做到设备的兼容性。

2. 独立的操作系统

Celerra的DART操作系统具备自主知识产权，专注于文件系统的传输。该操作系统功能强大，性能优越，保证了文件系统高速可靠的传输。 Celerra后端通过SAN网络连接后端存储设备，拥有多条链路冗余，避免单点故障，保障了数据安全性。用户的数据只要保存一个拷贝，即可被前端的各种类型的主机所使用，因此，具备主机无关性。Celerra的DART操作系统对于不同操作系统Unix和Windows同样保证了数据共享，并且各自的访问权限亦可得到相应的保证。

3. 安装及管理简便

NAS无需服务器直接上网，而是采用面向用户设计的、专门用于数据存储的简化操作系统，内置了与网络连接所需的协议，整个系统的管理和设置较为简单。Celerra只要现有的网络具有空闲的网口，在无需关机的情况下，即可提供给前端不同类型主机进行访问，无需在主机上安装任何的软硬件。

4. NAS底层协议

NAS采用了NFS（Sun）沟通Unix阵营和CIFS沟通NT阵营，这也反映了NAS是基于操作系统的“文件级”读写操作，访问请求是根据“文件句柄+偏移量”得出。

由于NAS以上诸多优点及良好的兼容性，笔者相信NAS未来将会得到更加广泛的应用。

【存储入门必读】SAN协议

原创 EMC中文技术社区 戴尔易安信技术支持 *2016-03-08*

SAN(Storage Area Network的简称)直译过来就是存储区域网络，它采用光纤通道(Fibre Channel)技术，通过光纤通道交换机连接存储阵列和服务器主机，建立专用于数据存储的区域网络。SAN网络存储是一种高速网络或子网络，SAN存储系统提供在计算机与存储系统之间的数据传输。一个SAN网络由负责网络连接的通信结构、负责组织连接的管理层、存储部件以及计算机系统构成，从而使SAN技术保证数据传输的安全性和力度。

SAN具有以下几点优势：

1. SAN的可扩展性意味着你有少数的磁盘不受连接到系统上的限制。SAN可以增长到数百个磁盘，但是普通物理服务器的极限只有十几个。

2. SAN的性能不会受到以太网流量或本地磁盘访问量的制约。数据通过SAN从自己的私有网络传送，隔开用户流量、备份流量和其他SAN流量。

3. 在正确的配置环境下，SAN数据被区域划分。用户保存数据的分区和其他人处在同样的SAN.SAN区域隔离就如同将UNIX服务器和Windows服务器连接到相同的SAN上，但这两种服务器上的数据访问是不同的，事实上，Windows系统不能“看到”UNIX的数据，反之亦然。

4. SAN系统不需要重新启动就能添加新的磁盘，更换磁盘或配置RAID组。数据流完全避开服务器系统，SAN同样增加了数据备份和恢复性能。

5. 分区也可以在SAN上将你的工作负载分离。不仅是将你的数据分离保护，而且对那些影响应用程序性能的不相关的工作负载采取屏蔽。在适当的区域应用SAN共享从性能上讲不是问题。

6. SAN有个无可比拟的优势，就是存储连接距离为10公里距离(约6英里)。不是说你一定会用到这个优势，但当你需要的时候，它就能显现出来。具有距离优势，可以将数据存储到一个独立的位置，从系统服务中脱离出来。

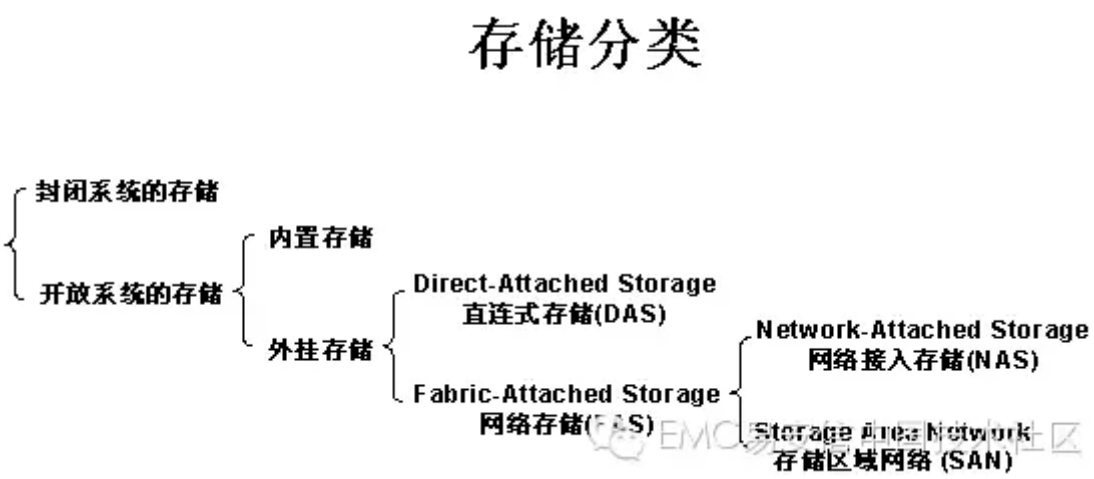
7. 在如SAN这样的存储网络上的自动精简配置的空间利用效率，要比本地存储的来得高。当一个系统需要更多的存储资源时，SAN将动态分配资源。这意味着物理系统可以享受自动精简配置，就像虚拟化那样。

【存储入门必读】NAS与SAN的区别

原创 EMC中文技术社区 戴尔易安信技术支持 *2016-03-09*

目前磁盘存储市场上，存储分类（如下表一）根据服务器类型分为：封闭系统的存储和开放系统的存储，封闭系统主要指大型机，AS400等服务器，开放系统指基于包括Windows、UNIX、Linux等操作系统的服务器；开放系统的存储分为：内置存储和外挂存储；开放系统的外挂存储根据连接的方式分为：直连式存储（Direct-Attached Storage，简称DAS）和网络化存储（Fabric-Attached Storage，简称FAS）；开放系统的网络化存储根据传输协议又分为：网络接入存储（Network-Attached Storage，简称NAS）和存储区域网络（Storage Area Network，简称SAN）。由于目前绝大部分用户采用的是开放系统，其外挂存储占有目前磁盘存储市场的70%以上，因此本文主要针对开放系统的外挂存储进行论述说明。

表一：

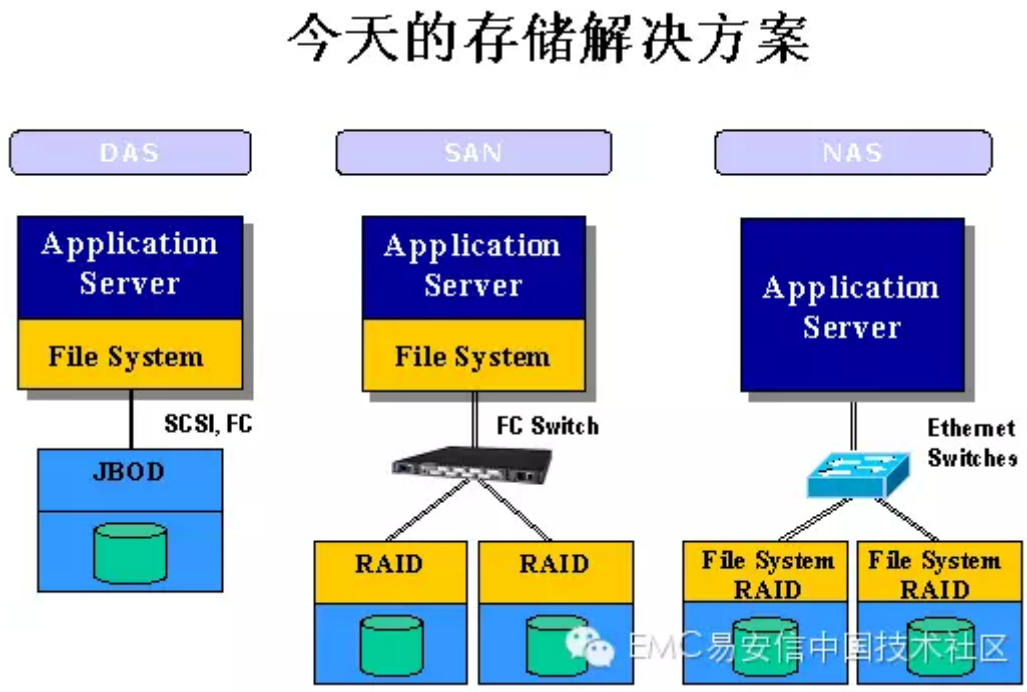


今天的存储解决方案主要为：直连式存储（DAS）、存储区域网络（SAN）、网络接入存储（NAS）。如下：

**

**

表二：



开放系统的直连式存储（Direct-Attached Storage，简称DAS）已经有近四十年的使用历史，随着用户数据的不断增长，尤其是数百GB以上时，其在备份、恢复、扩展、灾备等方面的问题变得日益困扰系统管理员。

主要问题和不足为：

**

**

直连式存储依赖服务器主机操作系统进行数据的IO读写和存储维护管理，数据备份和恢复要求占用服务器主机资源（包括CPU、系统IO等），数据流需要回流主机再到服务器连接着的磁带机（库），数据备份通常占用服务器主机资源20-30%，因此许多企业用户的日常数据备份常常在深夜或业务系统不繁忙时进行，以免影响正常业务系统的运行。直连式存储的数据量越大，备份和恢复的时间就越长，对服务器硬件的依赖性和影响就越大。

直连式存储与服务器主机之间的连接通道通常采用SCSI连接，带宽为10MB/s、20MB/s、40MB/s、80MB/s等，随着服务器CPU的处理能力越来越强，存储硬盘空间越来越大，阵列的硬盘数量越来越多，SCSI通道将会成为IO瓶颈；服务器主机SCSI ID资源有限，能够建立的SCSI通道连接有限。

无论直连式存储还是服务器主机的扩展，从一台服务器扩展为多台服务器组成的群集(Cluster)，或存储阵列容量的扩展，都会造成业务系统的停机，从而给企业带来经济损失，对于银行、电信、传媒等行业7×24小时服务的关键业务系统，这是不可接受的。并且直连式存储或服务器主机的升级扩展，只能由原设备厂商提供，往往受原设备厂商限制。

存储区域网络（Storage Area Network，简称SAN）采用光纤通道（Fibre Channel）技术，通过光纤通道交换机连接存储阵列和服务器主机，建立专用于数据存储的区域网络。SAN经过十多年历史的发展，已经相当成熟，成为业界的事实标准（但各个厂商的光纤交换技术不完全相同，其服务器和SAN存储有兼容性的要求）。SAN存储采用的带宽从100MB/s、200MB/s，发展到目前的1Gbps、2Gbps。

网络接入存储（Network-Attached Storage，简称NAS）采用网络（TCP/IP、ATM、FDDI）技术，通过网络交换机连接存储系统和服务器主机，建立专用于数据存储的存储私网。随着IP网络技术的发展，网络接入存储（NAS）技术发生质的飞跃。早期80年代末到90年代初的10Mbps带宽，网络接入存储作为文件服务器存储，性能受带宽影响；后来快速以太网（100Mbps）、VLAN虚网、Trunk(Ethernet Channel) 以太网通道的出现，网络接入存储的读写性能得到改善；1998年千兆以太网（1000Mbps）的出现和投入商用，为网络接入存储（NAS）带来质的变化和市场广泛认可。由于网络接入存储采用TCP/IP网络进行数据交换，TCP/IP是IT业界的标准协议，不同厂商的产品（服务器、交换机、NAS存储）只要满足协议标准就能够实现互连互通，无兼容性的要求；并且2002年万兆以太网（10000Mbps）的出现和投入商用，存储网络带宽将大大提高NAS存储的性能。NAS需求旺盛已经成为事实。首先NAS几乎继承了磁盘列阵的所有优点，可以将设备通过标准的网络拓扑结构连接，摆脱了服务器和异构化构架的桎梏。

其次，在企业数据量飞速膨胀中，SAN、大型磁带库、磁盘柜等产品虽然都是很好的存储解决方案，但他们那高贵的身份和复杂的操作是资金和技术实力有限的中小企业无论如何也不能接受的。NAS正是满足这种需求的产品，在解决足够的存储和扩展空间的同时，还提供极高的性价比。因此，无论是从适用性还是TCO的角度来说，NAS自然成为多数企业，尤其是大中小企业的最佳选择。

NAS与SAN的分析与比较

针对I/O是整个网络系统效率低下的瓶颈问题，专家们提出了许多种解决办法。其中抓住症结并经过实践检验为最有效的办法是：将数据从通用的应用服务器中分离出来以简化存储管理。

问题：

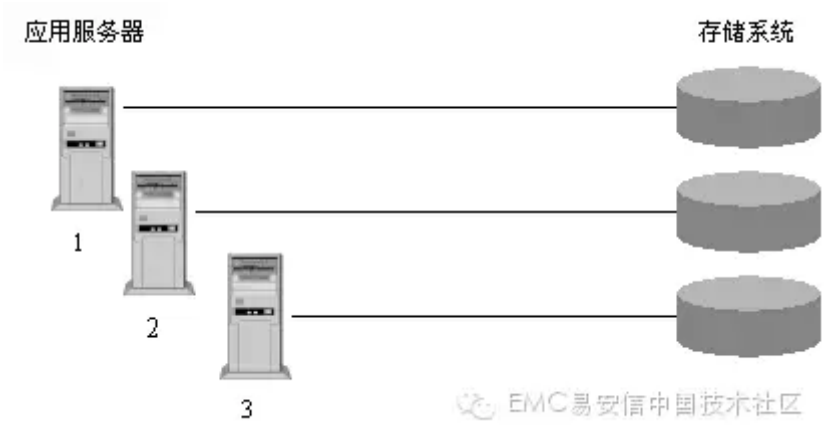


图 1

由图1可知原来存在的问题：每个新的应用服务器都要有它自己的存储器。这样造成数据处理复杂，随着应用服务器的不断增加，网络系统效率会急剧下降。

解决办法：

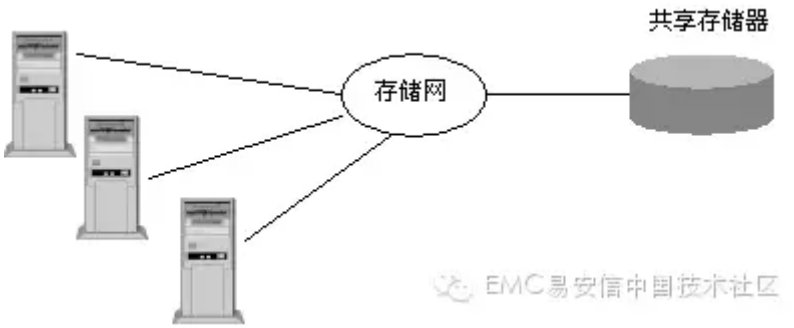


图 2

从图2可看出：将存储器从应用服务器中分离出来，进行集中管理。这就是所说的存储网络（Storage Networks）。

使用存储网络的好处：

统一性：形散神不散，在逻辑上是完全一体的。

实现数据集中管理，因为它们才是企业真正的命脉。

容易扩充，即收缩性很强。

具有容错功能，整个网络无单点故障。

专家们针对这一办法又采取了两种不同的实现手段，即NAS（Network Attached Storage）网络接入存储和SAN(Storage Area Networks)存储区域网络。

NAS：用户通过TCP/IP协议访问数据，采用业界标准文件共享协议如：NFS、HTTP、CIFS实现共享。

SAN：通过专用光纤通道交换机访问数据，采用SCSI、FC-AL接口。

什么是NAS和SAN的根本不同点?

NAS和SAN最本质的不同就是文件管理系统在哪里。如图：

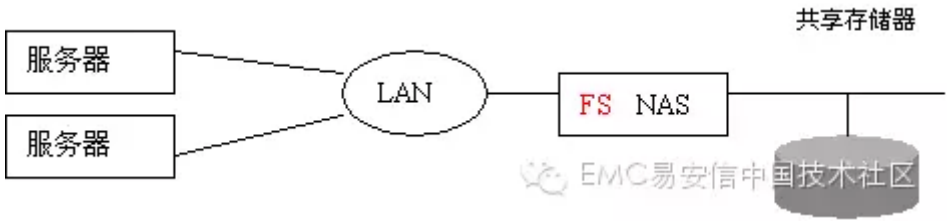


图3

由图3可以看出，SAN结构中，文件管理系统（FS）还是分别在每一个应用服务器上；而NAS则是每个应用服务器通过网络共享协议（如：NFS、CIFS）使用同一个文件管理系统。换句话说：NAS和SAN存储系统的区别是NAS有自己的文件系统管理。

NAS是将目光集中在应用、用户和文件以及它们共享的数据上。SAN是将目光集中在磁盘、磁带以及联接它们的可靠的基础结构。将来从桌面系统到数据集中管理到存储设备的全面解决方案将是NAS加SAN。

【存储入门必读】SCSI访问控制原理介绍

原创 EMC中文技术社区 戴尔易安信技术支持 *2016-03-10*

本文为大家介绍SCSI-2和SCSI-3访问控制原理。主要内容包括：SCSI-2 Reserve/Release/Reset和SCSI-3 Persistent Reserve IN/ Persistent Reserve OUT/ PREEMPT以及SCSI访问控制常见场景。

**
**

SCSI-2 Reserve(预留)/Release(释放)/Reset（重置）

SCSI-2协议中客户端访问lun过程如下：

1. 客户端向lun发起预留操作
2. 预留操作成功后，客户端获得lun操作权限；预留失败，提示预留冲突，会继续尝试，直到预留成功。
3. 客户端操作完毕后，执行释放操作，其他客户端可以预留。

SCSI-2访问控制主要缺点有：

1. 预留操作基于路径。预留和释放必须由相同的客户端完成，一台主机不能释放另外一台主机的预留，同一主机HBA卡不能取消相同主机另外一块HBA的预留。
2. 预留无法长久保留。主机重启将会丢失预留信息。
3. 如果lun已经被预留，其他主机无法再预留。如果其他主机要想获得lun操作权限，必须对lun进行重置，重置操作可能会导致数据丢失。重置后释放掉lun现有的预留，重置操作由lun主动发起，原来预留主机并不知晓。

SCSI-3 Persistent Reserve (PR)/ PREEMPT（抢占）

SCSI-3协议引入PGR（persistent group reservation）功能。在访问lun之前，客户端首先向lun注册（registration）一个预留密钥(reservation key)，注册成功后客户端可以尝试进行永久预留（reserve），永久预留成功后就可以获得lun操作权限。预留密钥是一串16进制的ASCII码，最长8个字节。永久预留一共6种类型，由1、3、5、6、7、8数字表示。包括两种操作类型和三种客户类型，操作类型包括写排它和所有访问排他，客户类型包括所有客户端、已注册客户端和所属客户端。数字与永久预留类型对应关系如下：

1-> write exclusive

3-> exclusive access

5-> write exclusive - registrants only

6-> exclusive access - registrants only

7-> write exclusive - all registrants

8-> exclusive access - all registrants.

不同注册类型对应不同访问权限。与SCSI-2不同，SCSI-3释放操作根据预留密钥。不同客户端可以使用相同密钥或是不同密钥进行预留，具体可以结合永久预留类型决定。客户端可以通过抢占来获取已被永久预留的lun访问权限。SCSI-3抢占和SCSI-2重置不一样，抢占不会造成数据丢失。

SCSI-3关于PGR相关操作命令分为两大类：分别是PRIN和PROUT。PRIN主要用于查询，PROUT用于修改。SCSI命令执行过程中，需要明确该命令是哪种类型。

常见使用场景

1. **集群I/O Fencing**

   为了防止集群故障发生“脑裂”现象，2-节点集群可以通过SCSI-2 Reseve/Release触发I/O fencing来保证整个集群正常运行，是SCSI-2不适用于多-节点集群，多-节点集群可以使用SCSI-3 PGR。主流厂商集群套件都已经支持SCSI-3 PGR，比如：VCS、HACAMP、RHCS等。

2. **集群文件系统**

   集群文件系统需要保证多节点同时访问存储时的数据一致性，SCSI-2/SCSI-3都可以满足，当一个节点尝试访问一个已经被预留的存储就会产生访问权限冲突。SCSI-3 PGR相比SCSI-2 Reserve/Release更能够减少访问权限冲突。

小结:

SCSI-2具体基本访问控制能力，但是无法满足Active/Active多路径环境和集群多节点访问存储的需求。SCSI-3通过引入客户端注册和操作权限分类概念，强化并行访问权限控制，弥补SCSI-2的不足。

NAS实现类型对比：统一式、网关式和横向扩展式（Scale-out）

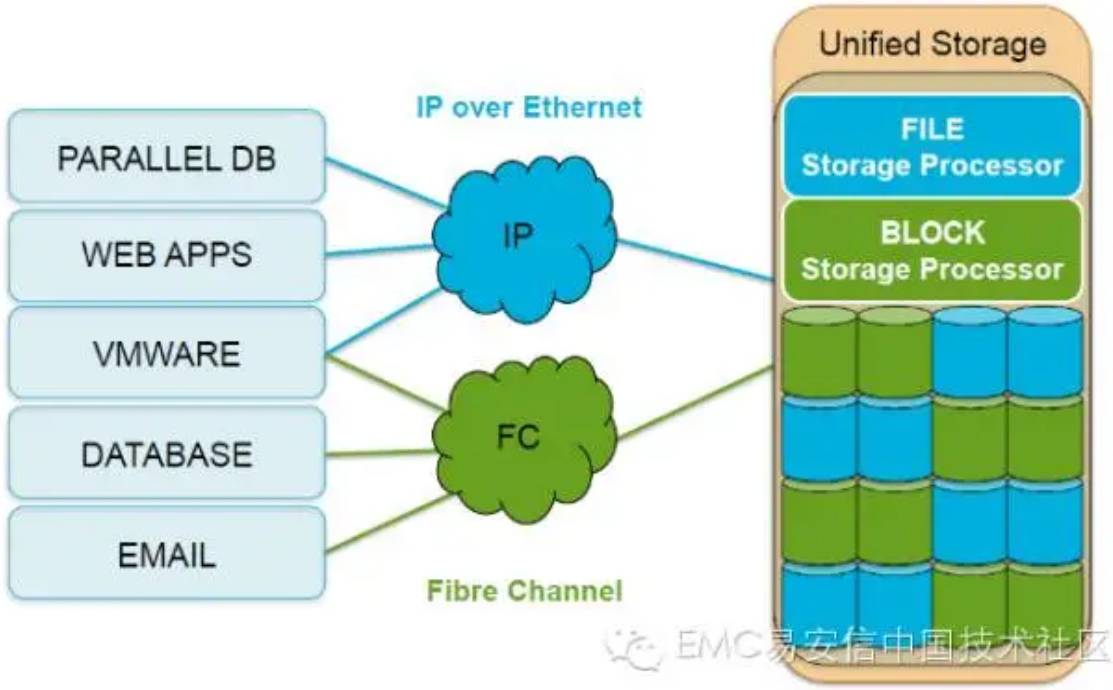
原创 EMC中文技术社区 戴尔易安信技术支持 *2016-03-04*

NAS主要有三种类型的实现：统一式、网关式和横向扩展式（Scale-out）。统一NAS使用统一的存储平台将基于NAS和基于SAN的数据访问合并，提供了可以同时管理二种环境的统一管理界面。网关NAS使用外部存储来存取数据，网关NAS和存储的管理操作是分开的。横向扩展式（Scale-out）NAS可组合多个节点，形成一个集群NAS系统。本文将对比三种不同NAS实现类型。

统一NAS

统一NAS提供文件服务，同时负责存储文件数据，并提供块数据访问。它支持用于文件访问的CIFS和NFS协议，以及用于块级访问的SCSI和FC协议。因为基于NAS和基于SAN的访问合并到同一个存储平台，统一NAS降低了企业的基础设施成本和管理成本。

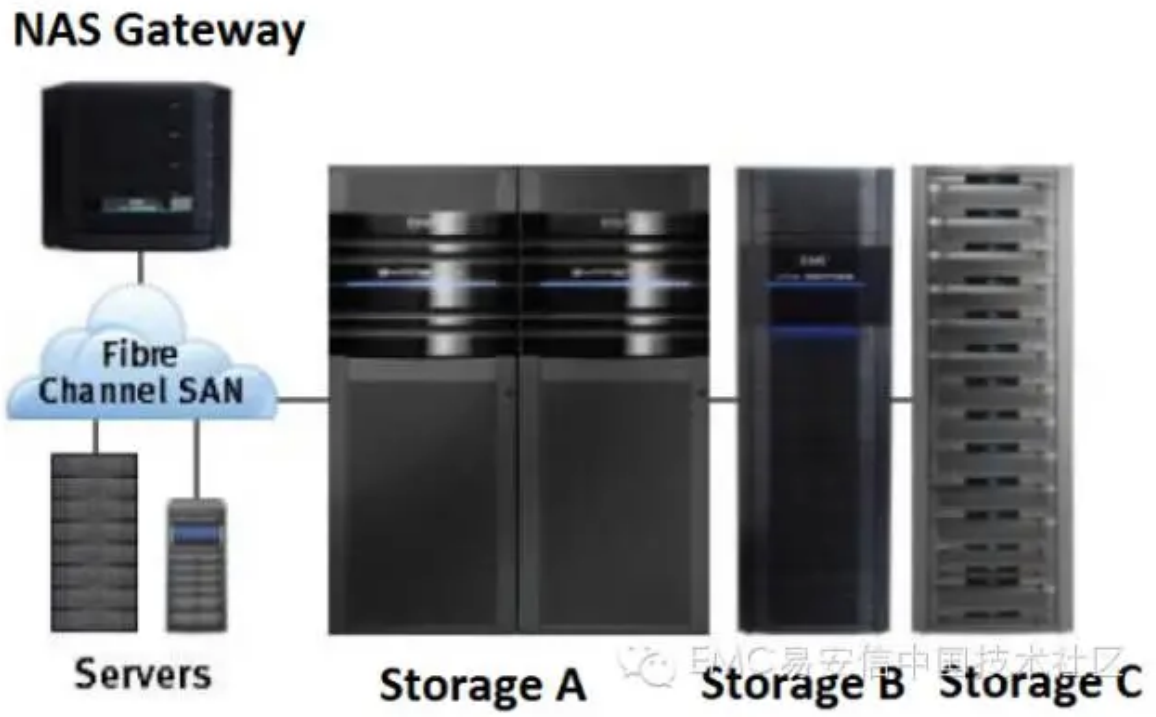
统一NAS的一个系统中包括了一个或多个NAS头及存储。NAS头与存储控制器连接，提供到存储的访问。存储控制器提供了与iSCSI和FC主机的连接。存储可使用不同的磁盘类型（例如SAS、ATA、FC和闪存盘），以满足不同的负载需求。下图显示的是一个统一NAS连接的例子。



网关式NAS

网关式NAS设备包含一个或多个NAS头，使用外部存储或者独立管理的存储。与统一NAS相似，存储是与其他使用块级I/O的应用共享的。这种解决方案的管理功能比统一存储复杂，因为NAS头和存储器的管理任务是分开的。网关式解决方案可以利用FC基础设施，例如：交换机等，访问SAN存储阵列或直连式存储阵列。

网关式NAS的扩展性比统一NAS好，因为NAS头和存储阵列可以独立地根据需求进行扩展升级。例如：可以通过增加NAS头的方式提升NAS设备的性能。当存储容量达到上限时，网关NAS设备可以独立于NAS头对SAN进行扩展，增加存储容量。网关式NAS通过在SAN环境中进行存储共享，提高了存储资源的利用率。下图显示的是一个网关式NAS连接的例子。



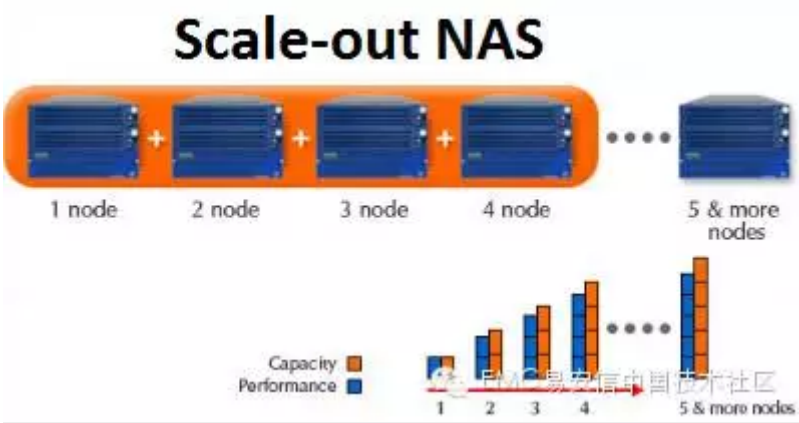
横向扩展式（Scale-out）NAS

统一NAS和网关NAS实现都提供了一定的扩展性能，可以在数据增长和性能需求提高时对资源进行扩展。对NAS设备进行扩展主要涉及增加CPU、内存和存储容量。扩展性受制于NAS设备对后续增加NAS头和存储容量的支持能力。

横向扩展式（Scale-out）NAS可组合多个节点，形成一个集群NAS系统。只需要向集群NAS架构中添加节点即可实现资源的扩展。整个集群可看作是一个NAS设备，资源是集中管理的。在需要扩大容量或提高性能的时候，可向集群中添加节点，这不会造成停机下线的情况。横向扩展NAS可以集合许多性能和可用性中等的节点，形成集群系统拥有更好的总体性能和可用性。它还有易使用、成本低以及理论上可无限扩展的优势。

横向扩展式（Scale-out）NAS在集群中的所有节点上创建了一个单一文件系统。节点的所有信息都可以彼此共享，因此连接到任何节点的客户端都可以访问整个文件系统。集群NAS将数据在所有节点间分条，同时使用镜像或效验方式进行数据保护。数据从客户端发送到集群时，数据被分割，并行分配给不同节点。当客户端发送文件读取请求时，集群NAS从多个节点获取相应的块，将他们组合成文件，然后将文件发给客户端。随着节点的增加，文件系统实现动态扩展，数据在节点之间均衡分布。每个增加的节点都增加了整个集群的存储、内存、CPU和网络能力。因此，整个集群的性能都得到提升。

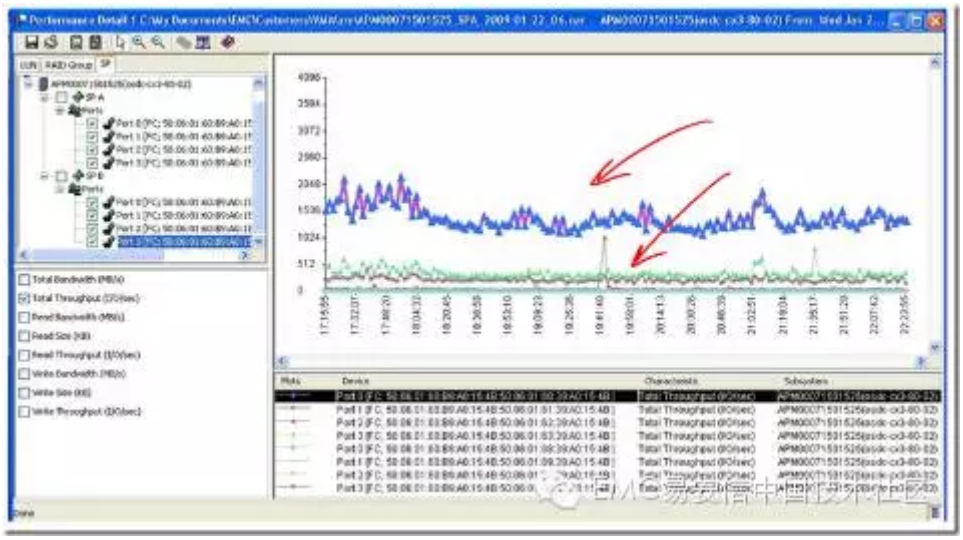
横向扩展式（Scale-out）NAS适合解决企业和客户当前面临的大数据问题。它统一管理和存储高速增长的数据，同时又十分灵活，能满足各种性能需求。下图显示的是一个横向扩展式（Scale-out）NAS连接的例子。



SAN网络性能问题排错指南

原创 EMC中文技术社区 戴尔易安信技术支持 *2016-03-21*

主机和FC阵列间出现了性能问题，应该如何排错？快来看看下面的建议，也许可以让你少走不少弯路。



1. 由于性能问题通常难以定位（特别当工程师在以远程的方式提供服务），事先收集足够详细和明确的信息可以节省后续大量的时间和精力

- 问题的详细描述
- 问题第一次出现是什么时候？是怎么发现的？
- 问题再次出现是什么时候？
- 问题最严重的时刻是何时？

- 出现了哪些症状?
- 主机上出现了哪些错误?
- 哪些设备（主机的LUN，大型机的UCB）受到影响?

2. 你是如何测量性能问题的?

- 曲线图
- 图表
- 使用的工具和监控协议（如SNMP）采样时间间隔

3. 最近针对SAN网络有什么变动？提供这些变动的细节

- SAN网络中增加或移除的设备，包括主机、存储阵列、远程复制设备和交换机
- 存放或备份的数据量的变化
- 整个网络带宽（SAN、LAN或WAN）的变化
- 任何其他会影响到性能的变化

4. 物理层是否已经都检查过了?

- 参考以下步骤排错光纤交换机端口通讯问题

  i. 确认涉及通讯故障的节点和交换机端口

  ii. 确认交换机端口状态是" Administratively Up"

  iii. 将SFP模块连同光纤线缆换到同一台交换机的其它插槽

  iv. 如果问题依旧，则交换机有问题。如果问题解决了，则可能是SFP模块、光纤线或节点HBA有问题

  v. 如果SFP模块、光纤线的问题都排除了，则继续检查主机端的HBA

- 如果需要，EMC工程师会帮忙开单并派遣现场工程师上门检查物理层设备
- 确认所有链路连通性节点（主机到交换机、远端阵列到交换机、光纤跳线板等）已尝试过物理复位（重插拔）

5. 提供出现性能问题的数据路径上端到端的设备信息

- 是否有同一数据链路上的其他设备也遇到了性能问题?
- 后端设备的型号是?
- 主机类型和版本是?
- HBA类型和版本是?
- 主机上安装的EMC软件和版本（比如PowerPath）
- 哪些应用受影响?

- 提供HBA的pWWN和目标阵列（FA/SP等）的pWWN
- 提供交换机物理接口信息
- 提供主机initiator连入的交换机日志
- 提供目标阵列连入的交换机日志
- 提供光纤网络逻辑图（Visio、网络管理软件、网络快照等）

EMC Connectivity售后团队会尽最大努力帮助客户分析并纠正性能问题。但售后团队主要的工作内容是故障修复（break/fix），如果这一问题最终被确认为不是因为故障而引起的性能问题，那我们会将问题移交给EMC Professional Service部门的性能分析团队来处理。如果您确实需要这一服务的话，EMC售后团队会帮您联络必要的资源。

存储性能瓶颈的成因、定位与排查

原创 EMC中文技术社区 戴尔易安信技术支持 *2016-03-15*

企业数据存储性能瓶颈常常会发生在端口，控制器和磁盘，难点在于找出引起拥塞的单元，往往需要应用多重工具以及丰富的经验来查找并解决。

本文详细阐述存储瓶颈发生最常见的四种情况，可能发生的拥塞点，需要监控的参数指标，以及部署存储系统的最佳实践。

**
**

数据存储瓶颈的四个常见场景：

以下是储瓶颈发生最常见的四种典型情况：

**
**

1.

当多个用户同时访问某一业务应用，无论是邮件服务器，企业资源规划（ERP）系统或数据库，数据请求会累积在队列中。单个I/O的响应时间开始增长，短暂延时开始转变成为漫长的等待。

这类响应时间敏感型应用的特征是，很多随机请求，读取比写入更多，I/O较小。最好的方法是：将负载分布在多块磁盘上，否则可能造成性能瓶颈。

如果应用增加了更多用户，或应用IOPS请求增加，则可能需要在RAID组中添加更多磁盘，或数据可能需要跨越更多磁盘，在更多层级做条带化。

存储在这样的情况下往往首先被怀疑，但大多数情况下并非存储引发，原因可能在于网络、应用或服务器。

2.

带宽敏感型应用——如数据备份，视频流或安全登录，这类应用当多个用户同时访问大型文件或数据流时可能造成瓶颈。

定位这一问题存储管理员应当从备份服务器开始一路向下检查至磁盘，原因可能存在于这一通路的任何地方。

问题不一定发生在存储，可能是由于备份应用创建的方式或是磁带系统的工作方式引起的。如果瓶颈定位于存储，那么可能是由于服务I/O的磁盘数量不足，在控制器造成争用，或是阵列前端口带宽不足。

性能调优需要针对不同应用程序负载来完成。针对大型文件和流数据的调优并不适合于小型文件，反之亦然。这也就是为什么在大多数存储系统中往往做一个平衡，需要用户尝试并找出系统的折中。用户通常需要优化吞吐量或IOPS，但并不需要对两者同时优化。

3.

RAID组中的磁盘故障。特别是在RAID 5中会造成性能的下降，因为系统需要重建校验数据。相比数据读写操作，重建会对性能造成更大影响。

即便坏盘是造成故障的根源，但控制器还是可能成为瓶颈，因为在重建过程中它需要不停地服务数据。当重建完成时，性能才会恢复正常。

4.

部署了一种新的应用，而卷存在于处理繁忙邮件系统的同一磁盘。如果新的应用变得繁忙，邮件系统性能将会遭受影响。额外的流量最终会将磁盘完全覆盖。

存储瓶颈常发区域:

存储区域网络（Storage-area network, SAN）/阵列前端口

存储部署于集中化SAN环境时，需考虑服务器和SAN之间的潜在网络瓶颈。例如，运行多部虚拟机的整合服务器可能不具备支持工作负载要求的足够网络端口。添加网络端口或转移网络密集型工作负载至其他服务器可解决这一问题。如前所述，对于带宽集中型应用，需考虑NFS有多少Fiber Channel 端口, or iSCSI 端口 or Ethernet 端口，需要用户站在带宽的角度来考量整个架构。

可能发生的问题包括：

- 如果阵列中端口数量不够，就会发生过饱和/过度使用。
- 虚拟服务器环境下的过量预定
- 端口间负载不均衡
- 交换机间链路争用/流量负荷过重
- 如某一HBA端口负载过重将导致HBA拥塞。使用虚拟机会导致问题更加严重。

存储控制器

一个标准的主动——被动或主动——主动控制器都有一个性能极限。接近这条上限取决于用户有多少块磁盘，因为每块磁盘的IOPS和吞吐量是固定的。

可能出现的问题包括：

- 控制器I/O过饱和，使得从缓存到阵列能够处理的IOPS受到限制
- 吞吐量"淹没"处理器
- CPU过载/处理器功率不足
- 性能无法跟上SSD

Cache

由于服务器内存和CPU远比机械磁盘快得多，需为磁盘添加高速内存以缓存读写数据。例如，写入磁盘的数据存储在缓存中直到磁盘能够跟上，同时磁盘中的读数据放入缓存中直到能被主机读取。Cache比磁盘快1000倍，因此将数据写入和读出Cache对性能影响巨大。智能缓存算法能够预测你需要查找的数据，你是否会对此数据频繁访问，甚至是将访问频繁的随机数据放在缓存中。

可能发生的问题包括：

- Cache memory不足
- Cache写入过载，引起性能降低
- 频繁访问顺序性数据引起cache超负荷
- Cache中需要持续不断地写入新数据，因此如果cache总是在refill，将无法从cache获益。

磁盘

磁盘瓶颈与磁盘转速有关, 慢速磁盘会引入较多延时。存储性能问题的排查首先考虑的因素就是磁盘速度，同时有多少块磁盘可进行并发读写。而另一因素是磁盘接口。采用更快的接口能够缓解磁盘瓶颈，但更重要的是在快速接口与相应更大的缓存大小以及转速之间取得平衡。同样，应避免将快速和慢速磁盘混入同一接口，因为慢速磁盘将会造成快速接口与快速磁盘的性能浪费。

可能引发的问题包括：

- 过多应用命中磁盘
- 磁盘数量不足以满足应用所需的IOPS或吞吐量
- 磁盘速度过慢无法满足性能需求及支持繁重工作负荷
- Disk group往往是classic存储架构的潜在性能瓶颈，这种结构下RAID最多配置在16块磁盘。Thin结构通常每个LUN拥有更多磁盘，从而数据分布于更多spindle，因增加的并发性而减少了成为瓶颈的可能。

需要监控的指标：

曾经一度存储厂商们强调的是IOPS和吞吐量，但现在重点逐渐转变成为响应时间。也就是说，不是数据移动的速度有多快，而在于对请求的响应速度有多快。

正常情况下，15,000 rpm Fibre Channel磁盘响应时间为4ms，SAS磁盘响应时间约为5ms至6ms，SATA为10ms，而SSD少于1ms。如果发现Fibre Channel磁盘响应时间为12ms，或SSD响应时间变成5ms，那么就说明可能产生了争用，可能芯片发生了故障。

除了响应时间，其他需要监控的指标包括：

- 队列长度，队列中一次积累的请求数量，平均磁盘队列长度；
- 平均I/O大小千字节数；
- IOPS （读和写，随机和顺序，整体平均IOPS）；
- 每秒百万字节吞吐量；
- 读写所占比例；
- 容量（空闲，使用和保留）。

数据存储性能最佳实践：

性能调优和改进的方式有很多种，用户当然可以通过添加磁盘，端口，多核处理器，内存来改善，但问题是：性价比，以及对业务是否实用。本文建议的方式是在预算范围内找寻性能最大化的解决方案。另外一个需要考虑的方面是环境并非一尘不变，系统部署方案要能够适应环境的改变需求。

首先需要考虑刷数据的性能特征，需要了解IO工作情况是怎样的。是否是cache友好型？是否是CPU集中型？业务数据很大数量很少，还是很小但数量很多？另外一方面就是构成存储环境的组件。包括应用，存储系统本身，网络。。。瓶颈可能在哪里，改善哪里最有效？

以下是一些常规建议：

1. 不要仅仅根据空闲空间来分配存储，而需要结合考虑性能需求，确保为吞吐量或IOPS分配足够多的磁盘。
2. 在磁盘间均衡分布应用负载，以减少热点地区的产生。
3. 理解应用负载类型，并针对负载选择匹配的RAID类型。例如，写密集型应用建议使用RAID 1而不是RAID 5。因为当写入RAID 5时，需要计算校验位，需耗费较多时间。而RAID 1，写入两块磁盘速度快得多，无需计算。
4. 磁盘类型（Fibre Channel, SAS, SATA）与期望性能相匹配。对于关键业务应用部署高性能磁盘，例如15,000 rpm Fibre Channel。
5. 对于I/O密集型应用考虑采用SSD，但并不适用于写性能重要型应用。只要没有达到控制器瓶颈，SSD对读性能提升显著，但对写性能提升并没有明显效果。
6. 采用端对端的监控工具，特别是虚拟服务器环境。虚拟端与物理端之间有一道防火墙，所以，需要穿透防火墙进行端到端的监控。
7. 有些性能分析工具涵盖从应用到磁盘，有些仅局限于存储系统本身。由于性能是一个连锁反应包含很多变量，所以需要全面地分析数据。
8. 以数据仅写入磁盘外部扇区的方式格式化磁盘。因减少数据定位时间而在高I/O环境下提升性能。负面作用是相当一部分磁盘容量未能得以使用。

什么是zone？如何做zone？如何做好zone？

原创 EMC中文技术社区 戴尔易安信技术支持 *2016-03-17*

一．什么是zone

Zone是FC-SAN交换机上的一种独有的逻辑配置，通过配置特定的设备加入zone，从而允许设备之间互相通信。当交换机上配置了zone时，同在一个zone里的设备之间可以互相通信，没有加入任何zone的设备不能与其他设备通信。

早期交换机厂商根据zone的实现方式，把zone分为hard zone和soft zone，区别在于前者通过硬件芯片来实现，后者通过软件来实现。后来大家把基于domain ID/端口号的zone叫做hard zone，基于wwn的zone为soft zone。现在这两种类型的zone都是基于硬件芯片实现。

Zone的类型：

**

**

1. 基于Domain ID/端口号(D,P)的普通zone模式

这种zone允许接在某几个端口上的设备互相通信，即使端口上的设备改变也不会影响zone的使用，在更换主机HBA卡时不需要进行任何zone配置的更改。

2. 基于wwpn/wwnn的普通zone模式

这种zone允许拥有特定wwn的设备之间互相通信，不关心设备接在交换机的哪个口上。当某个设备从一个端口移到另一个端口时，不需要进行任何zone配置的更改。但更换主机HBA卡时，需要根据新HBA卡的wwn更改zone配置。注意如果交换机上接有NPIV模式的刀片交换机或主机集群时，必须使用基于wwn的zone。

3. 混合zone(session based hard zoning)

当一台设备在两个或多个zone里分别使用D,P和wwn模式的zone，这台设备会进入混合zone模式。在混合zone模式里的设备在跟其他设备通信时需要通过交换机CPU进行软件验证。

4. LSAN zone

LSAN zone只有在启用了FCR时才会被应用到，它能允许在不同的fabric中的设备通过fc router进行通信。需要在交换机上安装integrated routing license后才能打开FCR功能。

5. TI zone(Traffic Isolation zone)

TI zone可以把一根或者多根ISL设置成某个zone的专用ISL，不需要license。

6. QOS zone

QOS zone在网络中出现拥堵时可以允许高QOS的zone成员优先通信，需要在交换机上安装adaptive networking license。

Zoneset是zone的集合。一台交换机同时只能启用一个zoneset，同一个SAN网络中交换机的active zoneset必须保持一致，不然会造成网络分裂(fabric segment)。

Alias，或叫做别名，是使配置zone更简便的一个功能。对于每台设备，可以预先设置好alias，之后在配置zone时使用alias来代替D,P或wwn。

Default zone：思科与博科交换机都有default zone，它的功能是在没有任何zone配置时允许所有连接在交换机上的设备互相通信。

二．如何做zone

博科交换机CLI命令行：

首先对每个需要做zone的设备创建alias，然后创建zone并把alias加入，创建cfg(zoneset)并把需要的zone加入，最后启用cfg。

帮助命令： zonehelp

显示现有配置：cfgshow

创建/增加成员/移除成员/删除alias：

alicreate "aliName","member[; member...]"

aliadd "aliName","member[; member...]"

aliremove "aliName","member[; member...]"

alidelete "aliName"

创建/增加成员/移除成员/删除zone:

zonecreate "zonename", "member[;member...]"

zoneadd "zoneName", "member[;member...]"

zoneremove "zoneName", "member[;member...]"

zonedelete "zoneName"

注意：根据zone的最佳实践，EMC推荐每个zone里只放一个initiator(主机，Vplex的BE口等)。多个initiator互相zone在一起会导致很多反常现象。

创建/增加成员/移除成员/删除cfg：

cfgcreate "cfgName", "member[;member...]"

cfgadd "cfgName", "member[;member...]"

cfgremove "cfgName", "member[;member...]"

cfgdelete "cfgName", "member[;member...]"

保存/启用cfg：

cfgsave

cfgenable "cfgName"

注意：激活某个cfg会使其他正被使用cfg停止工作，一个fabric里同时只能有一个cfg处于工作状态。

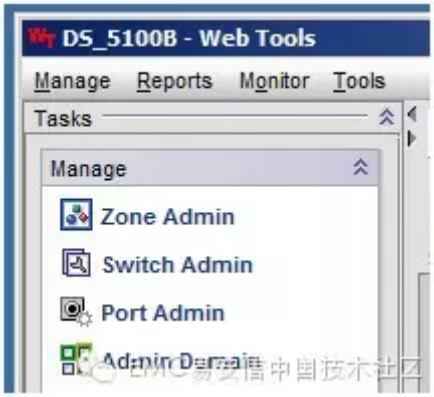
更改default zone配置：

defzone [--noaccess | --allaccess | --show]

博科交换机GUI界面：

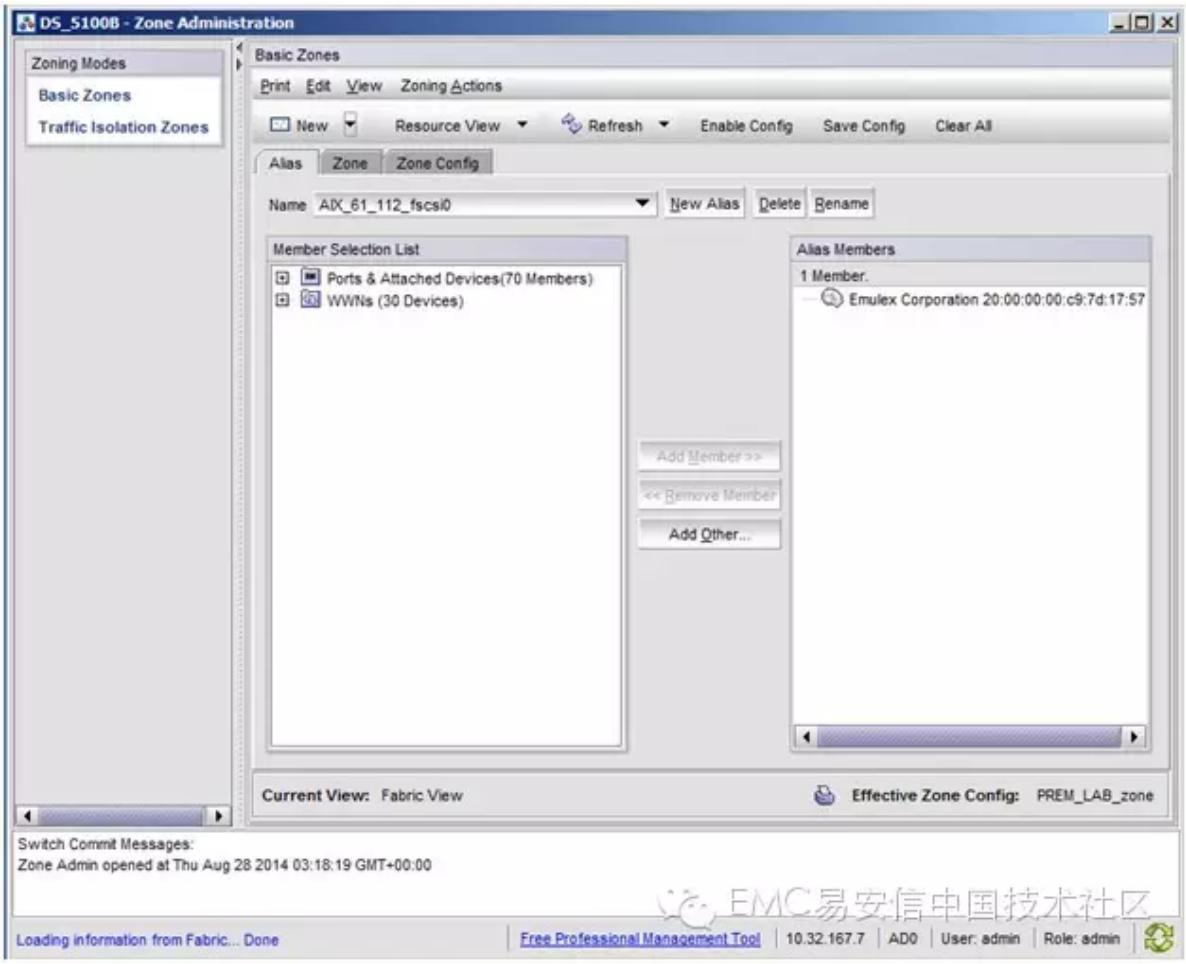
进入webtools后点击Zone Admin，进入zone配置界面。

V6.x.x界面：

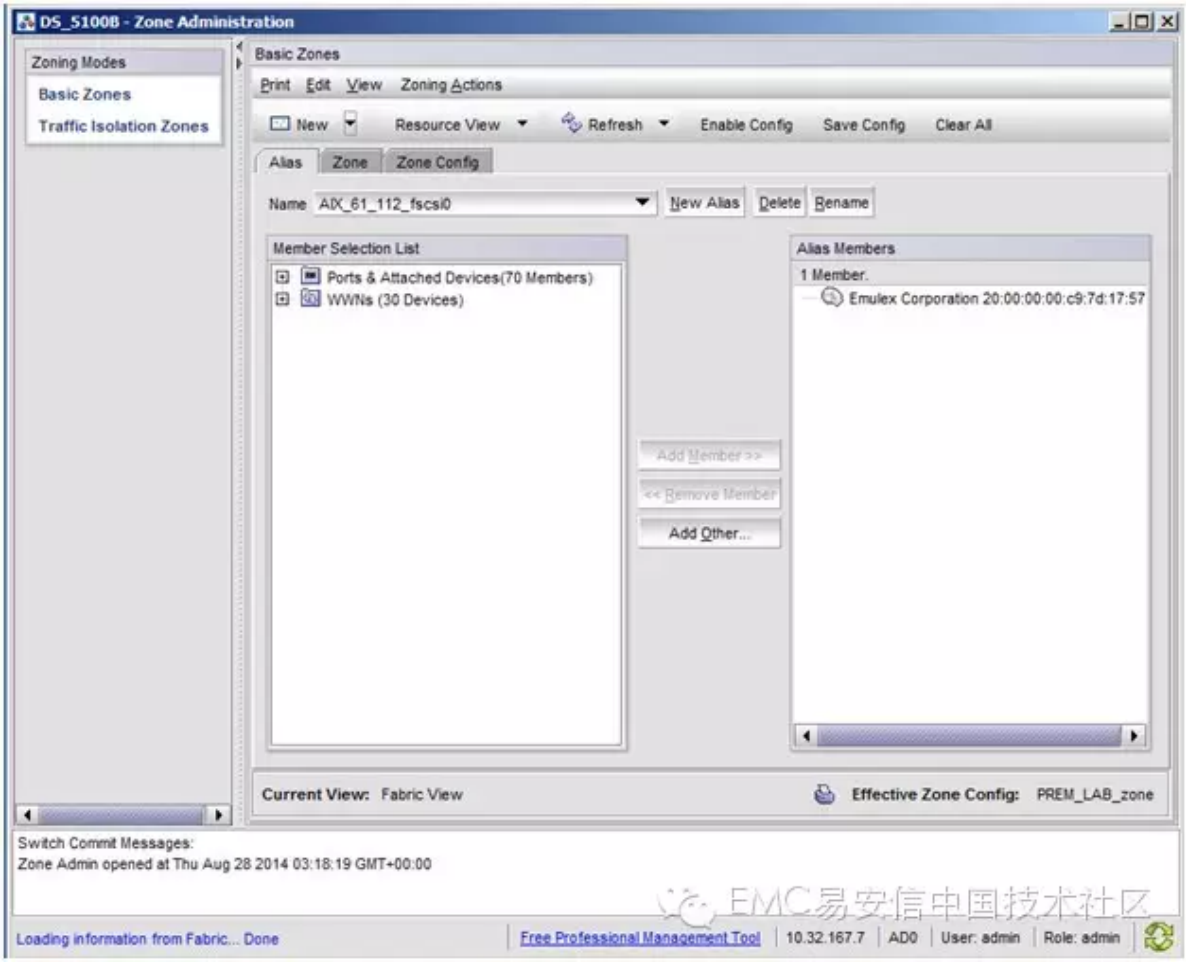


V7.x.x界面：

进入Zone Admin后v6.x.x与v7.x.x版本的界面基本一致。



创建alias：

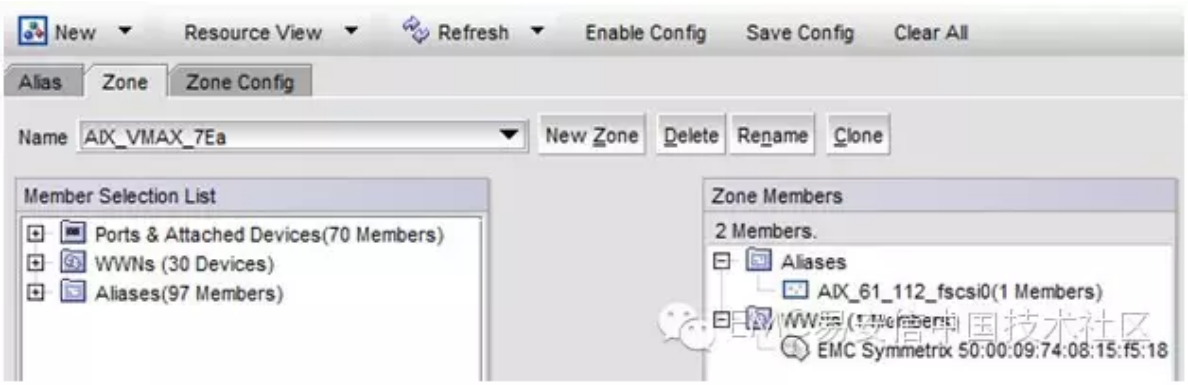


点击New键或右边的new alias键，输入alias名字，注意只能输入数字字母或下划线。

点击OK后注意  栏内已经显示刚才输入的alias名字，然后从左边的列表里选中相应的wwn或交换机端口，点击add member键加入右边的alias members里。

创建zone并添加成员：

选中标签页中的zone标签，点击New按键，输入zone名字并点击OK。

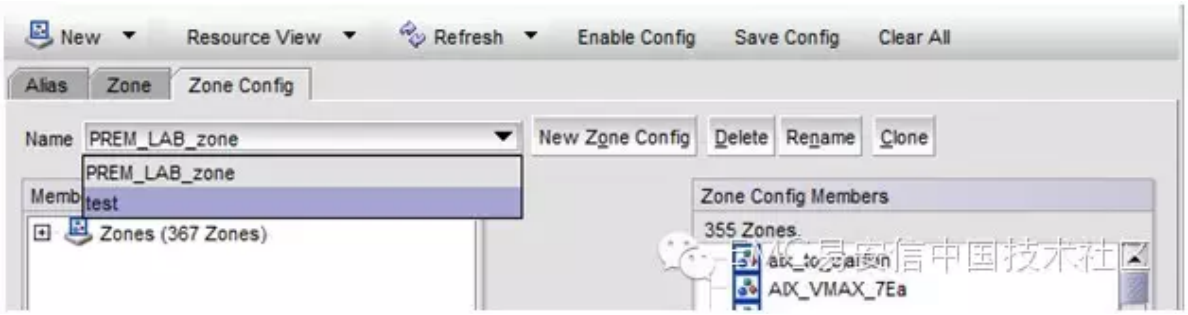


然后从左边列表里选中相应的wwn，交换机端口或之前设置好的alias，点击add member键加入右边的zone members里。

注意：根据zone的最佳实践，EMC推荐每个zone里只放一个initiator(主机，Vplex的BE口等)。多个initiator互相zone在一起会导致很多反常现象。

创建cfg并添加成员：

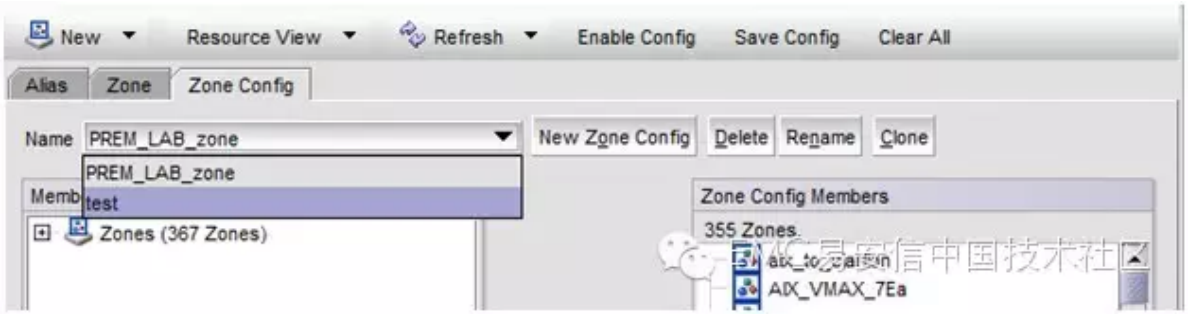
选中标签页中的zone config标签，点击New键，输入cfg的名字并点击OK。



然后从左边列表里选中相应的zone，点击add member键加入到右边的zone config members里。

保存并激活cfg：

选中标签页中的zone config标签，查看name右边下拉菜单，确认当前的cfg是需要激活的cfg。



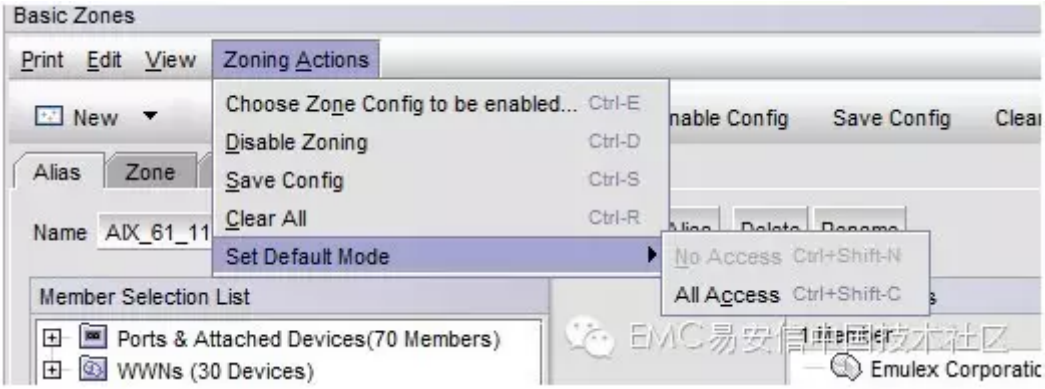
点击save config按钮，保存之前更改好的cfg。

点击enable config按钮，激活当前选中的cfg。

注意：激活某个cfg会使其他正被使用cfg停止工作，一个fabric里同时只能有一个cfg处于工作状态。

更改default zone配置：

点选zoning actions菜单，选中set default mode里的no access或all access。



思科交换机CLI命令行：

思科交换机与博科交换机最大的不同就是vsan，每个vsan都拥有自己独立的zone和zoneset。

其次还有enhanced zoning与basic zoning的区别。

enhanced zoning会在用户试图更改zone配置时创建一个session，防止其他用户同时更改配置造成配置丢失。开启了enhanced zoning功能的交换机在做完zone配置更改之后需要commit以使配置生效并关闭session。

另外需要注意的是enhanced zoning会自动开启广播zone，而MDS9500系列在升级到第四代端口板的时候需要禁用广播zone才能是第四代端口板生效。

显示命令：

```
# show fcalias vsan x

# show zoneset vsan x
```

show active zoneset vsan x

show zone status vsan x

启用enhanced zoning：

configure terminal

(config)# zone mode enhanced vsan x

更改alias：

(config)# fcalias name A123 vsan x

(config-fcalias)# member pwwn 10:00:00:00:00:00:00:00

(config-fcalias)# exit

(config)# zone commit vsan x

更改zone：

(config)# zone name zone123 vsan x

(config-zone)# member interface fc1/1

(config-zone)# member pwwn 20:00:00:00:00:00:00:00

(config-zone)# member fcalias A123

(config-zone)# exit

(config)# zone commit vsan x

更改zoneset：

(config)# zoneset name zoneset123 vsan x

(config-zoneset)# member zone123

(config-zoneset)# exit

(config)# zone commit vsan x

激活zoneset(只在basic zone模式下有效)：

(config)# zoneset activate name zoneset123 vsan 1

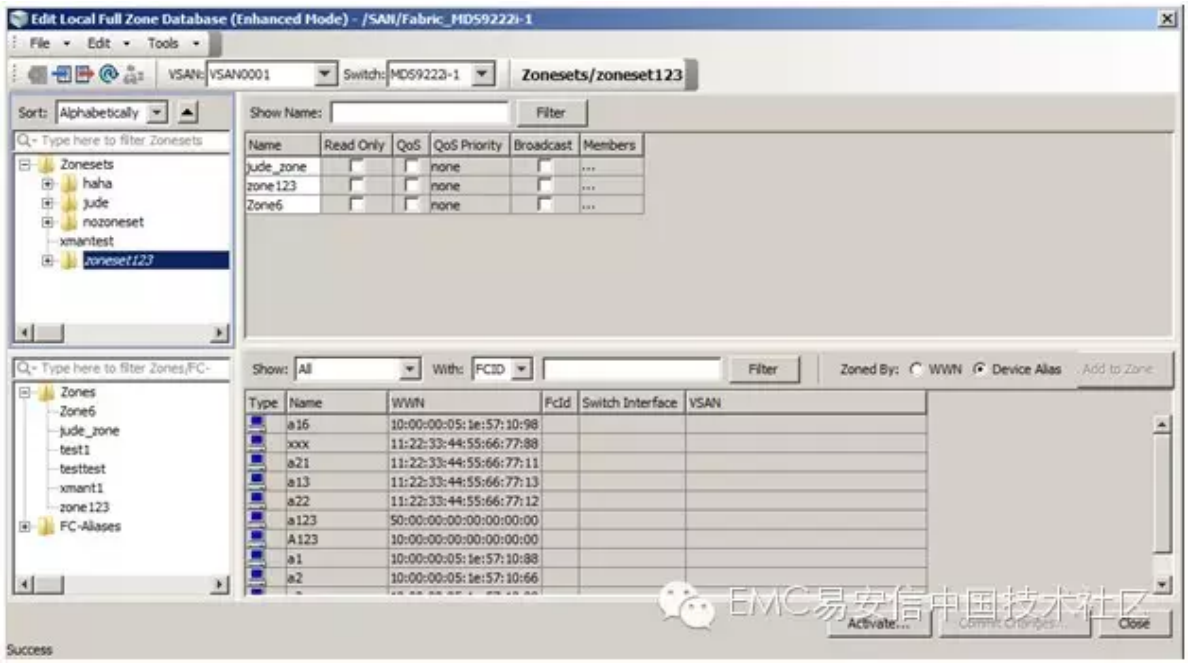
禁用广播zone：

(config)# no zone broadcast enable vsan x

思科交换机GUI界面(DCNM与DCFM基本一致)：

点击DCNM界面zone菜单中的edit local full zone database...

编辑zone的界面如下：

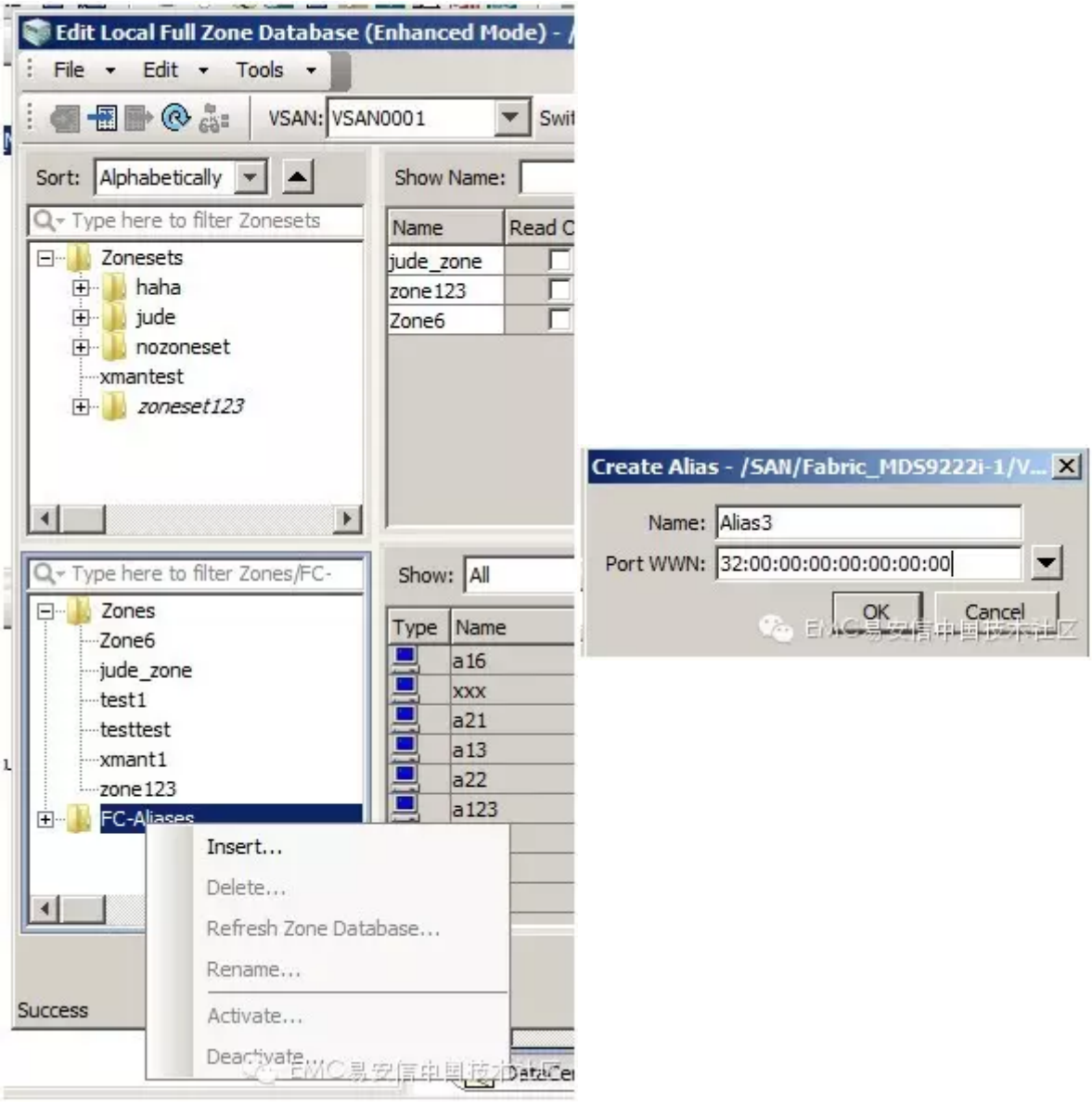


**

**

编辑fcalias：

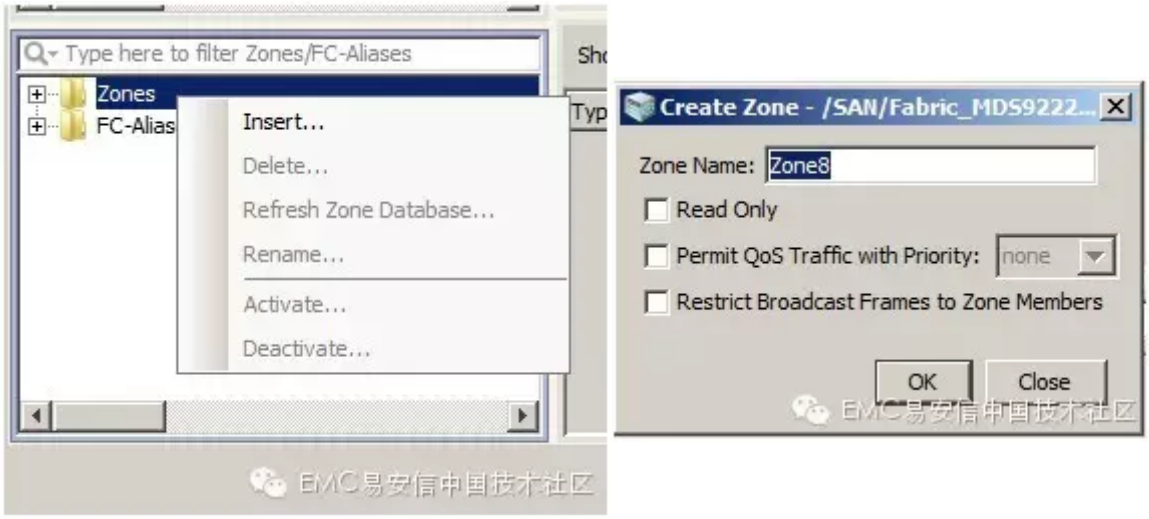
首先在左下角找到fc-alias，右键点击insert并输入alias名和wwn：



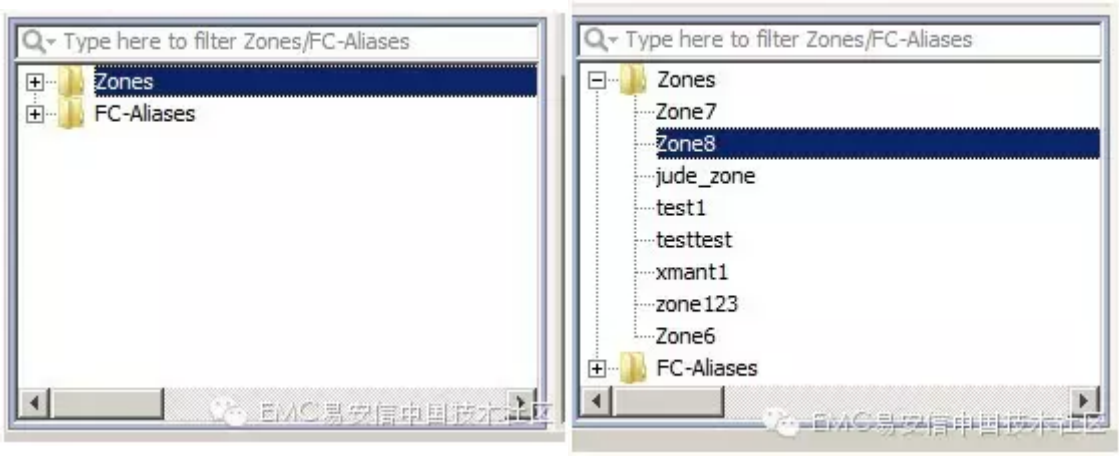
点击OK后创建alias的窗口并不会立刻关闭，可以更改alias名和wwn，再点击OK来连续创建其他alias。

编辑zone：

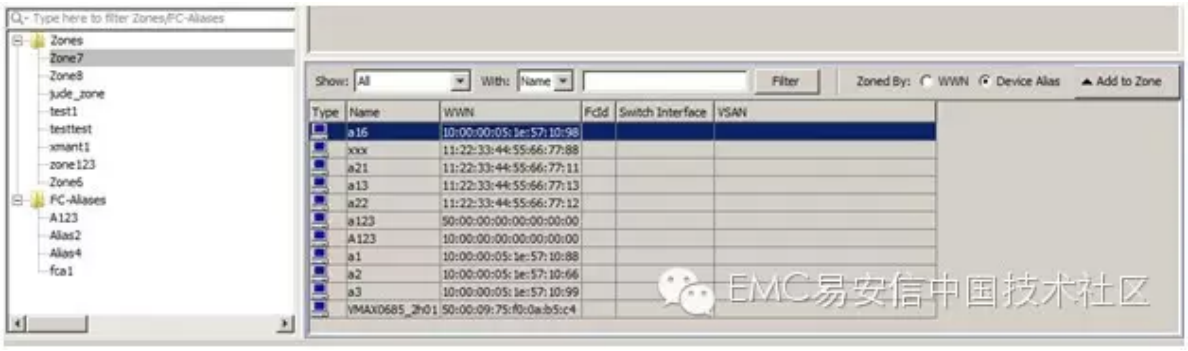
点击左下角的zones，右击并选中insert，可以创建zone。



点击OK后会看到新创的zone显示在列表里，在左下角点开zones前的加号，选中新增的zone来编辑其成员。



在右下角的列表里找到相应的wwn或device alias，点击add to zone加入到zone里，或在左下角把fcalias拖进相应的zone里。

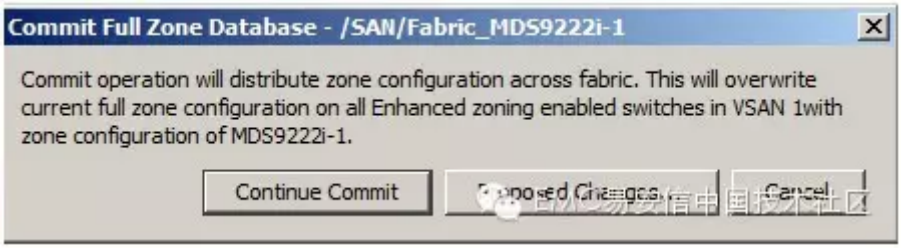


编辑zoneset：

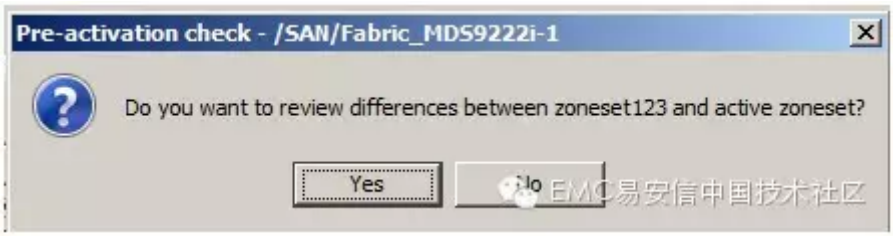
把之前编辑好的zone从左下角拖进左上角的zoneset即可。

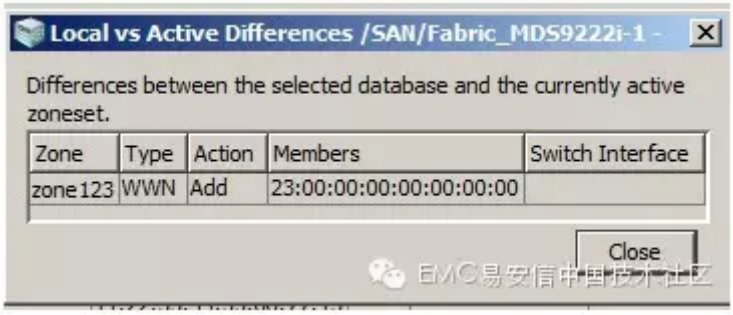
确认编辑/激活zoneset：

点击右下角的commit changes按钮，会把对该vsan做的更改发布到整个SAN网络里。

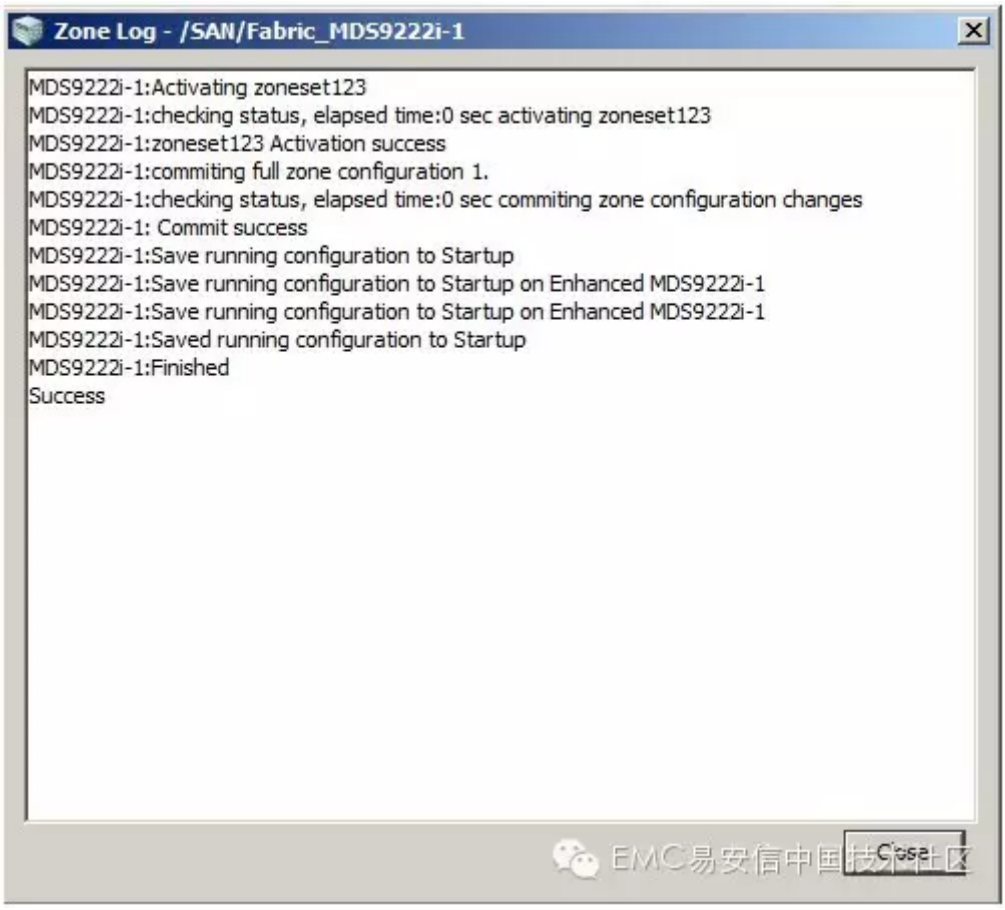


选中某个zoneset，点击右下角的activate按钮，会显示之前对这个zoneset做过的更改。





点击close之后显示保存running-config到startup-config的对话框，如果确实要执行该操作，打上选项前的勾并点击continue activation。



显示success了就说明激活zoneset并保存配置成功了。

三．如何做好zone

做一个zone很简单，但是如何做好zone，却要考虑到方方面面的问题。

1. 推荐使用wwn zone(客户有特殊要求或FICON环境除外)，原因如下：
    1. port zone只能通过物理隔离来保证zone安全，而wwn zone能限制只有指定设备才能访问zone。
    2. NPIV和AG环境中，只能使用wwn zone来划分zone给cluster上的主机或虚机。
    3. IVR/FCR和磁带加速技术只能使用wwn zone。
2. LUN masking和zone同时使用

    Zone和LUN masking都可以隔离主机和存储之间的通信，但是这两者作用在不同的层面。Zone在交换机上面生效，LUN masking在存储端口生效，两者无法做相互取代。
3. alias命名应该清晰易懂，确保不会混淆。
4. 博科交换机尽量避免使用混合zone模式。博科交换机在6.4.3之前有一个bug，会导致在混合zone里的主机自动登出存储。
5. 思科交换机使用enhanced zoning，防止多个用户同时更改zone配置导致配置丢失。
6. 关闭default zone，避免未经验证的设备登入网络。

CIFS安全协议之Kerberos

原创 EMC中文技术社区 戴尔易安信技术支持 *2016-03-16*

大家都知道对于NAS的CIFS 来说，NTLM和Kerberos是两个比较重要的验证方式。今天我们来讨论一下作为CIFS安全认证之一的Kerberos协议。

Kerberos 是一种依赖于验证技术共享密钥的协议，其基本概念很简单，如果一个秘密只有两个人知道，任何一个人都可以通过他们之间共享的秘密来确定对方的身份。用技术的语言来讲，就是对称加密，互相确认。

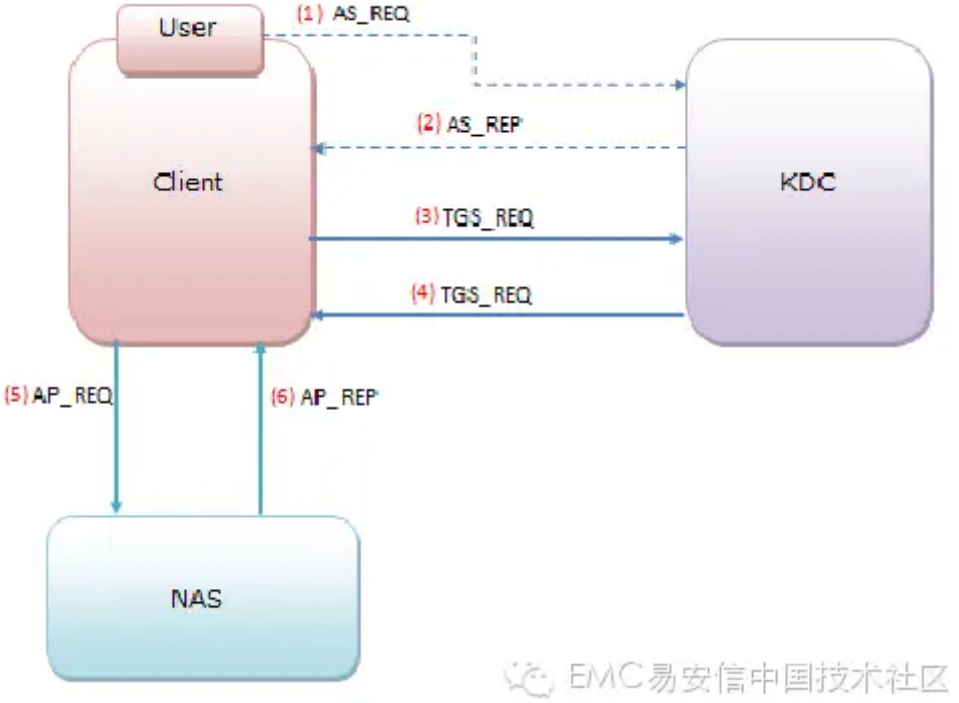
说到这里，可能有人会问，NTLM也是CIFS的安全认证啊，为什么Kerberos会用的更广泛呢？他的优点又在哪呢？相信大家了解了Kerberos的工作原理之后就会清楚了。

Kerberos认证和我们看电影的过程差不多，主要分三个步骤：

第一步咨询：用户登陆domain

第二步买票：用户获得service票据

第三步来访：使用服务票据访问某服务



用户登陆domain (Logging into the Domain)

\1. Authentication Server request

AS_REQ，即上图步骤(1)

一个用户在一台Client机上第一次登陆时，会有弹框提示输入用户名和密码。这时用户密码信息会通过Hash算法产生一个用户的Long Term Key(LTK)，再和用户登陆Client端时的时间戳一起进行加密，然后这个用户验证请求被发送给Kerberos Data Center(KDC)，并要求KDC返回一个相应的Ticket Granting
Ticket(TGT)。

\2. Authentication Server response

AS_REP，即上图步骤(2)

KDC在数据库中先找到该用户的密码，并用同样的Hash算法生成一个LTK。 然后KDC通过LTK从预认证信息Preauth包KRB_AR_REQ中解密出用户信息和时间戳， 如果用户信息无误，并且时间戳和目前信息相差在5分钟之内，KDC会认为该用户验证通过。KDC将生成一个TGT，并准备把TGT返回给Client。

在生成TGT的同时，KDC生成一组随机数作为Logon Session Key，并让他和客户端传来的LTK再次进行加密，产生一个名叫enc-part的信息放在该KRB_AS_REP包中，KDC还会用生成的TGT与能被所有KDC识别的LTK进行加密，再次产生一个enc-part放在TGT的头中，这样就得到了一个只有KDC能解开的TGT，最后把这个TGT加入KRB_AS_REP包中并发送至Client端。

\3. Ticket cache

Client端收到KRB_AS_REP后会用自己的LTK来解密KRB_AS_REP中的enc-part，这里我们默认会得到KDC选用的随机数Logon Session Key。这时候Client端将会把得到的Logon Session Key和KRB_AS_REP中带有原始时间戳的TGT一起存入票据内存中准备给下面过程使用。

PS：这时候KDC为了减少自身负担，并没有吧Logon Session Key保存在自己这边，所有只有这个用户在Client端的票据内存里面才有该记录了。

用户获得service票据(Getting a Service Ticket)

\1. Ticket Granting Server request

TGS_REQ，即上图步骤(3)

Client端接下来会拿着TGT去告诉KDC我要访问某服务，并让KDC给他访问该服务的票据(service ticket)，以后他就可以拿着这张票据直接访问该服务了. Client端用现在的时间戳和之前存在票据内存中的Logon Session Key进行加密生成Authenticator，然后再把带有原始时间戳的TGT一起打包发送给KDC等待验证。

\2. Ticket Granting Server response

TGS_REP，即上图步骤(4)

KDC收到KRB_TGS_REQ后会用自己的LTK解密出TGT，然后看看TGT中的原始时间戳，如果来确定TGT现在还是处于有效的状态，KDC就会从TGT中读取Logon Session Key，并用这个Logon Session Key解密Authenticator来获得第二次记录的时间戳，如果该时间差在5分钟之内，KDC认为验证成功，并将生成一个service ticket。

PS：Kerberos为了防止重演攻击，特别加入了对时间戳的验证。

接下来KDC又会生成一组叫做Service Session Key的随机数，并把这组随机数和service ticket中记录的Logon Session Key进行加密，产生一个名叫enc-part的信息放在该KRB_TGS_REP包中，KDC还会用生成的service ticket与能被所有KDC识别的LTK进行加密，再次产生一个enc-part放在TGS的头中，这样就得到了一个只有KDC能解开的service ticket，最后把这个service ticket加入KRB_AS_REP包中并发送至Client端。

\3. Ticket Cache

Client收到KRB_TGS_REP后，通过票据内存中的Logon Session Key解密enc-part，然后得到KDC生成的Service Session Key。最后把Service Session Key和service ticket一起作为访问目标服务器的Credential存入票据内存中。

使用服务票据访问某服务(Using the Service Ticket)

\1. Application Server request

AP_REQ，即上图步骤(5)

经过上面的两次验证，Client现在就能拿着服务票据去访问该服务了。Client记录下时间戳，并读取内存中的Service Session Key与他进行加密生成新的Authenticator，同时用户还要标记好自己是否需要双向认证 (Mutual
Authentication)，最后再加上之前的服务票据一起发送给该服务的server(在NAS系统中相对应的服务就是CIFS，server就是我们加入domain的CIFS server)。

\2. Application Server response

AP_REP，即上图步骤(6)

NAS收到这个加密的KBR_AP_REQ之后，用自己的LTK进行解密。接着server从service ticket里面读出service session Key来解密Authenticator中的时间戳。如果时间差小于5分钟，NAS就允许该用户对他进行访问了，并且为这个Client上的这个用户创建一个security token。

Kerberos优点：

虽然上面Kerberos的工作原理稍微有点复杂，不过我们还是能从中看出Kerberos的高效性，相互身份验证以及互操作性的优点。

1. 高效性：客户端不用每次访问NAS时都去DC验证，而通过查询client credentials就可以验证了。
2. 相互身份验证：client和server可以互相验证。
3. 互操作性：微软的Kerberos V5实现是基于IETF的推荐标准规范。这样，Windows Server 2003的Kerberos V5实现就为其他使用Kerberos V5协议的网络的互操作打下了基。