存储基础（一）

注：本文转载自 EMC 中国研究院官方轻微博 http://qing.weibo.com/emclabschina，特此鸣谢。

EMC 中国研究院云基础架构实验室高级研究员 万林涛
EMC 中国研究院云应用平台实验室高级研究员 张 芸

本文简要介绍了存储领域的若干重要术语，旨在帮助大家能更好地了解、学习存储这一领域。限于作者个人水平、精力有限，如有不当之处敬请多多包涵。

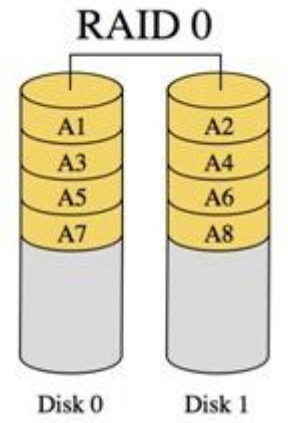
1. DAS (Direct-attached Storage)

直连式存储，顾名思义这是一种通过总线适配器直接将硬盘等存储介质连接到主机上的存储方式，在存储设备和主机之间没有任何网络设备的参与，此概念主要用于 区别 NAS 和 SAN 等网络存储。可以说 DAS 是最原始最基本的存储方式，在个人电脑、服务器上随处可见。常见的用于连接 DAS 和主机系统的协议/标准主要 有 ATA、SATA、eSATA、SCSI、SAS 和 FibreChannel 等。DAS 的优势在于简单易用、读写效率高等；缺点在于容量有限、难于共 享，从而造成“信息孤岛”（Islandsof Information）。

2. RAID (Redundant Array of Independent/InexpensiveDisks)

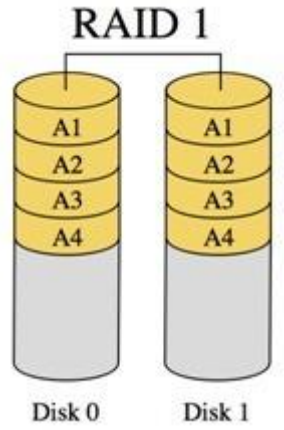
独立磁盘冗余阵列，是一种将多块独立的硬盘（物理硬盘）按不同的组合方式形成一个硬盘组（逻辑硬盘），从而提供比单块硬盘更大的存储容量、更高的可靠性和 更快的读写性能等。该概念最早由加州大学伯克利分校的几名教授于 1987 年提出。早期主要通过 RAID 控制器等硬件来实现 RAID 磁盘阵列，后来出现了基 于软件实现的 RAID，比如 mdadm 等。按照磁盘阵列的不同组合方式，可以将 RAID 分为不同级别，包括 RAID0 到 RAID6 等 7 个基本级别，以及 RAID0+1 和 RAID10 等扩展级别。不同 RAID 级别代表着不同的存储性能、数据安全性和存储成本等。下面我们将分别介绍这几种 RAID 级别。

RAID 0： 简单地说，RAID0 主要通过将多块硬盘“串联”起来，从而形成一个更大容量的逻辑硬盘。RAID0 通过“条带化（striping）”将数据分成不同的 数据块，并依次将这些数据块写到不同的硬盘上。因为数据分布在不同的硬盘上，所以数据吞吐量得到大大提升。但是，很容易看出 RAID0 没有任何数据冗余， 因此其可靠性不高。

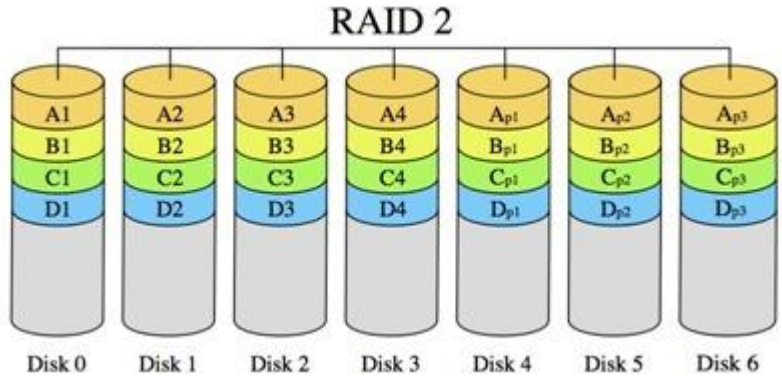


RAID 1：如果说 RAID 0 是 RAID 中一种只注重存储容量而没有任何容错的极端形式，那么 RAID1 则是有充分容错而不关心存储利用率的另一种极端表现。RAID1 通过“镜像 （mirroring）”，将每一份数据都同时写到多块硬盘（一般是两块）上去，从而实现了数据的完全备份。因此，RAID1 支持“热替

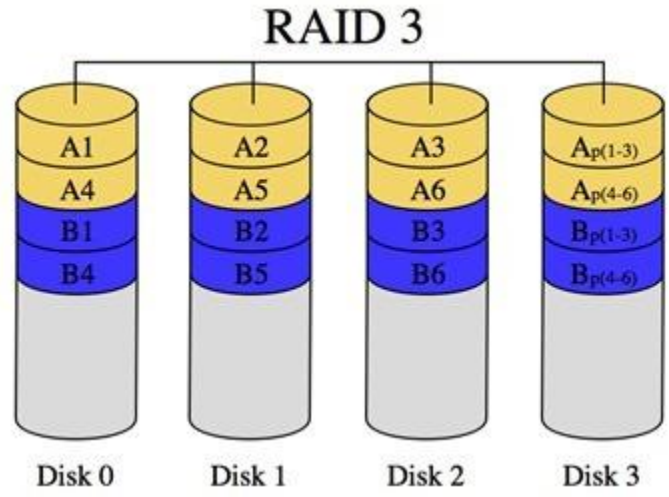
换”，在不断电的情况下对故障磁盘进行更换。一般情况下，RAID1 控制器在读取数据时支持负载平衡，允许数据从不同磁盘上同时读取，从而提高数据的读取速度；但是，RAID1 在写数据的性能没有改善。



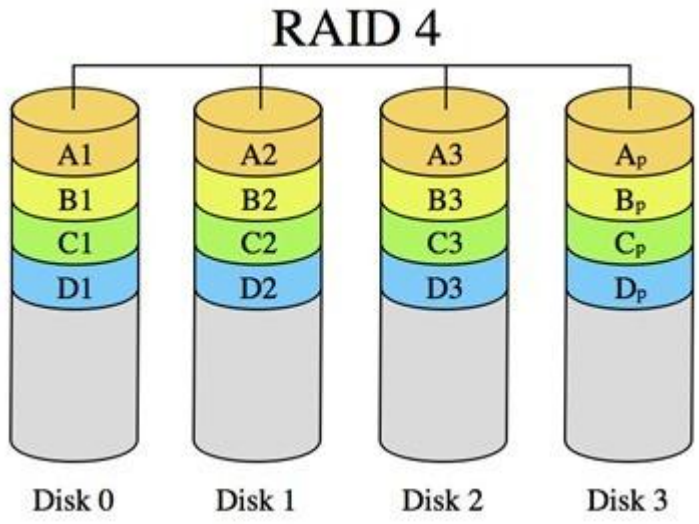
RADI 2：RAID 2 以比特（bit）为单位，将数据“条带化（striping）”分布存储在不同硬盘上；同时，将不同硬盘上同一位置的数据位用海明码进行编码，并将这些 编码数据保存在另外一些硬盘的相同位置上，从而实现错误检查和恢复。因为技术实施上的复杂性，商业环境中很少采用 RAID2。



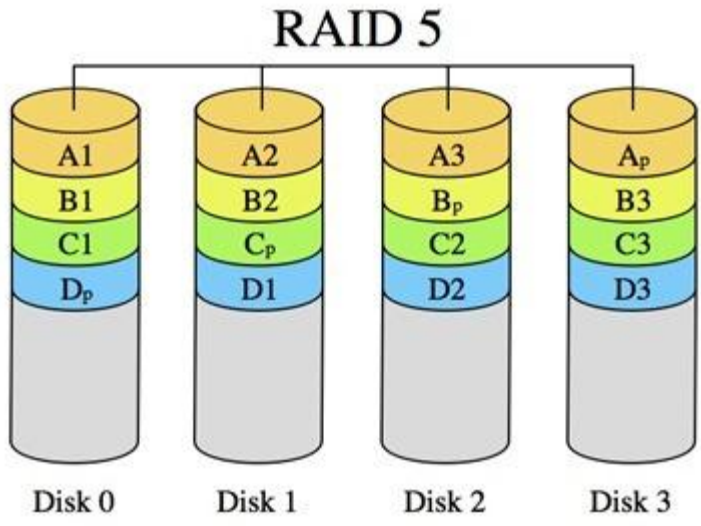
RAID 3：与 RAID 2 类似，不同的是：1）以字节（byte）为单位进行“条带化”处理；2）以奇偶校验码取代海明码。RAID3 的读写性能都还不错，而且存储利用率也相当高，可达到(n-1)/n。但是对于随即读写操作，奇偶盘会成为写操作的瓶颈。



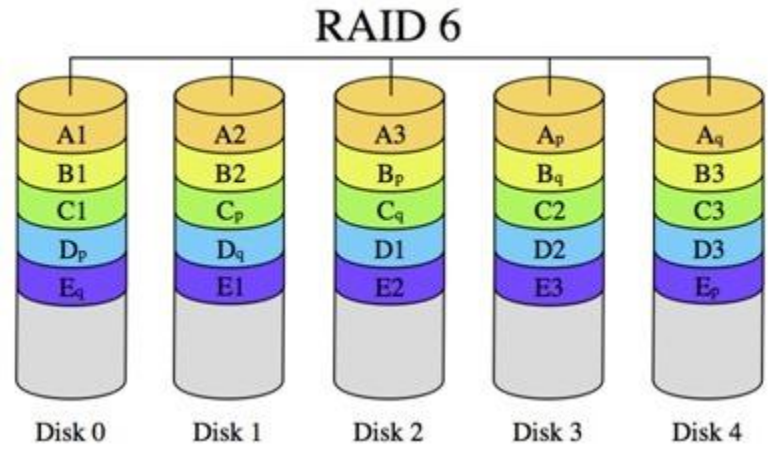
RAID 4：与 RAID 3 的分布结构类似，不同的是 RAID 4 以数据块（block）为单位进行奇偶校验码的计算。另外，与 RAID2 和 RAID3 不同的是，RAID4 中各个磁盘是独立操作的，并不要求各个磁盘的磁头同步转动。因此，RAID4 允许多个 I/O 请求并行处理。



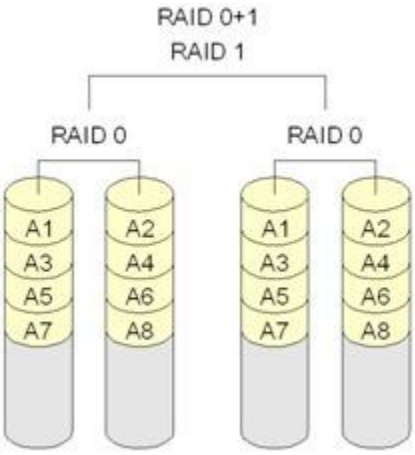
RAID 5: RAID 3 和 RAID 4 都存在同一个问题，就是奇偶校验码放在同一个硬盘上，容易造成写操作的瓶颈。RAID5 与 RAID4 基本相同，但是其将奇偶校验码分开存放到不同的硬盘上去，从而减少了写奇偶校验码带来瓶颈的可能性。



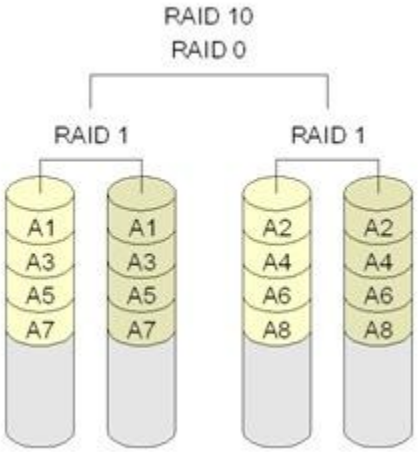
RAID 6：在 RAID 5 的基础上，RAID 6 又另外增加了一组奇偶校验码，从而获得更高的容错性，最多允许同时有两块硬盘出现故障。但是，新增加的奇偶校验计算同时也带来了写操作性能上的损耗。



RAID 0+1: 为了获取更好的 I/O 吞吐率或者可靠性，将不同的 RAID 标准级别混合产生的组合方式叫做嵌套式 RAID，或者混合 RAID。RAID0+1 是先将硬盘分 为若干组，每组以 RAID0 的方式组成“条带化”的硬盘阵列，然后将这些组 RAID0 的硬盘阵列以 RAID1 的方式组成一个大的硬盘阵列。



RAID 10: 类似于 RAID 0+1， RAID 10 则是先“镜像”（RAID 1）、后“条带化”（RAID0）。RAID0+1 和 RAID10 性能上并无太大区别，但是 RAID10 在可靠性上要好于 RAID0+1。这是因为在 RAID10 中，任何一块硬盘出现故障不会影响到整个磁盘阵列，即整个系统仍将以 RAID10 的方式运行；而 RAID0+1 中，一个硬盘出现故障则会导致 其所在的 RAID0 子阵列全部无法正常工作，从而影响到整个 RAID0+1 磁盘阵列 – 在只有两组 RAID0 子阵列的情况下，整个系统将完全降级为 RAID0 级别。



3. JBOD （Just a Bunch Of Disks）

JBOD 是指在一个底板上安装的带有多个磁盘驱动器的存储设备，类似于 RAID，但实质上又不同。因此中文多将其翻译为磁盘簇。JBOD 通过 SCSI 电缆 或其他连接方式，将多块磁盘串联起来，组成一个大的逻辑磁盘 。因此，相比起一堆松散的磁盘，JBOD 更容易管理。典型的 JBOD 磁盘系统可以容纳 8 个或者 16 个硬盘。连接在这个 JBOD 上的服务器将所有这些磁盘识 别成为一系列单独的磁盘。因此，一个 16 个硬盘合并的 JBOD 磁盘系统需要 16 个设备地址。在一些 I/O 技术，比如 SCSI 和光纤环路中，这会引发设备地 址不足等问题。

和 RAID 磁盘阵列不同，JBOD 没有前段逻辑来管理磁盘上的数据分布，每个磁盘都是进行独立寻址。常见的情况是，JBOD 上的数据简单地从第一块磁盘开 始存储，当第块磁盘存储空间用完后，在依次存储到其他磁盘上。另外，相比起智能磁盘系统，JBOD 磁盘系统尤其不适合于支持 RAID 或者其他类型的虚拟 化。尽管如此，如果的确需要的话，这些都可以在 JBOD 磁盘系统的外部实现，比如使用服务器端的软件实现或者在存储网络中采用单独的虚拟化实体。相比较于 RAID 磁盘阵列，JBOD 缺乏数据安全保障和数据分布管理等功能，但是基于其低成本、大容量和易管理等特点，JBOD 最近几年在业界还是得到了广泛的使 用。

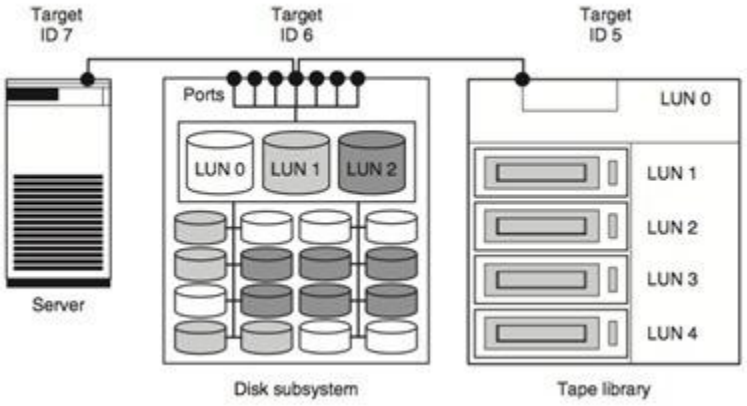
4. SCSI (Small Computer System Interface)

小型计算机系统接口，是一种智能的通用接口标准，定义了一系列用于连接计算机和各种外部设备的命令、协议以及接口规范等，常见用于硬盘和磁带等设备。第一 个版本的 SCSI 标准于 1986 年由美国国家标准协会（ANSI）制定，至今已形成很多新的标准和扩展等，比如 FastSCSI、UltraSCSI 等。

SCSI initiator 和 **target**: 在一个 SCSI 会话（session）中，负责发起会话和发送 SCSI 命令的一端被称作 initiator。而另一端主要负责接收、处理各种 SCSI 命 令，并负责数据的传输，被称作 target。简单地讲，可以分别将 initiator 和 target 类比于 C/S 架构中的客户端（client）和服务端 （server）。一般情况下，用户计算机或服务器扮演 initiator 的角色，而存储设备承担了 target 的角色。

SCSI ID 和 **LUN**: 依照不同版本的 SCSI 标准，一个 SCSI 总线最多可以连接 8 个或 16 个 SCSI 设备—实际情况中，在总线的末端一般要安装一个 SCSI 终结器 （terminator），所以最多可用的 SCSI 设备为 7 个或 15 个。每个连接在 SCSI 总线上的设备都一个唯一的 ID 号。鉴于一个 SCSI 总线上设备 数量的限制（最多 8 个或 16 个），一般 SCSI 存储设备都会由若干个子设备组成，比如 RAID 磁盘阵列、磁带

库等。为了标识这些子设备，SCSI 标准引入 了 LUN 的概念，即逻辑单元号（LogicalUnit Number）。所以，一个 SCSI 会话中，为了标识一个 SCSItarget，需要同时指明 SCSI 控制器 ID、SCSI ID 和 LUN。



SCSI 有多种接口形式，早期主要是并行的 SCSI 接口（如 SPI）；在 2008 年，串行的 SCSI 接口 SAS（Serial Attached SCSI）以其在数据吞吐率方面的潜力和优势取代了 SPI。另外，还有 SCSI 和光纤信道结合产生的 FCP 协议（SCSI-over- FibreChannel Protocol）。而 ISCSI 则完全摒弃 SCSI 在接口上规范，将 SCSI 的命令和协议等其他部分移植到 TCP/IP 网络上来，通过普通的 TCP /IP 网络来传输 SCSI 的数据和命令等。

5. SAN

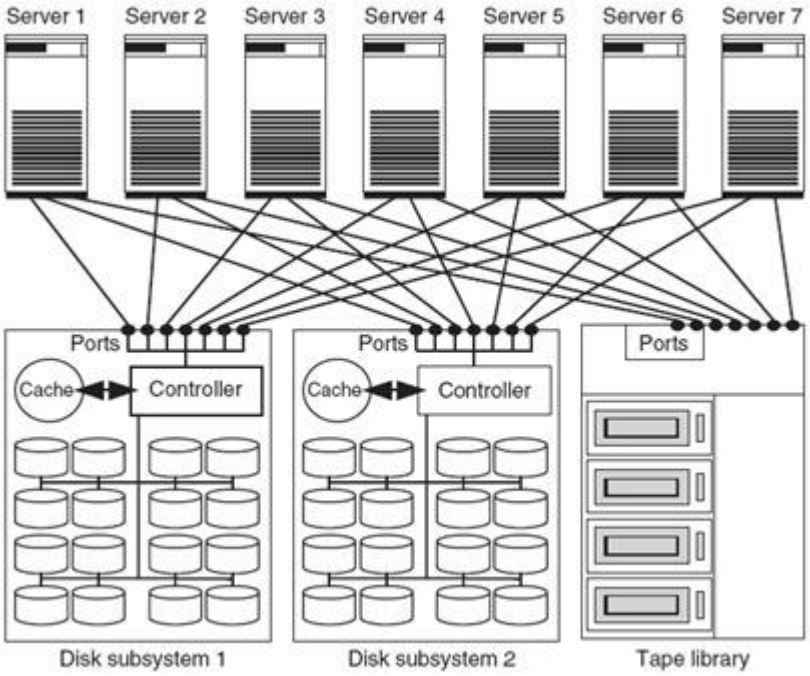
SAN 的全称是 Storage Area Network，是指一个提供统一的，块级别数据存取的专门网络。SAN 主要用来制作存储设备，比如磁盘阵列，磁带库等，服务器可以读取这些设备，就好像这些设备是操作系统的一个逻辑上附加的设备。

一个典型的 SAN 有它自己存储设备的网络，这个网络通常不能被其他设备的本地局域网访问到。SAN 只提供块级别的操作，而并不提供文件级别的操作。但是， 文件系统可以建立在 SAN 之上，以提供文件级别的读取，这种文件系统被称为 SAN 文件系统或者共享磁盘文件系统。

EMC 的 CLARiiON CX4 系列，就是这样一种提供块级别数据存取的 SAN 存储。

SAN 通常和 Fibre Channel（光纤通道）技术一同使用，表示采用光纤通道建立的一块存储区域网络，称为 FCSAN。同时，SAN 还可以和其他技术一起使用，比如 iSCSISAN 或者 FCoESAN。

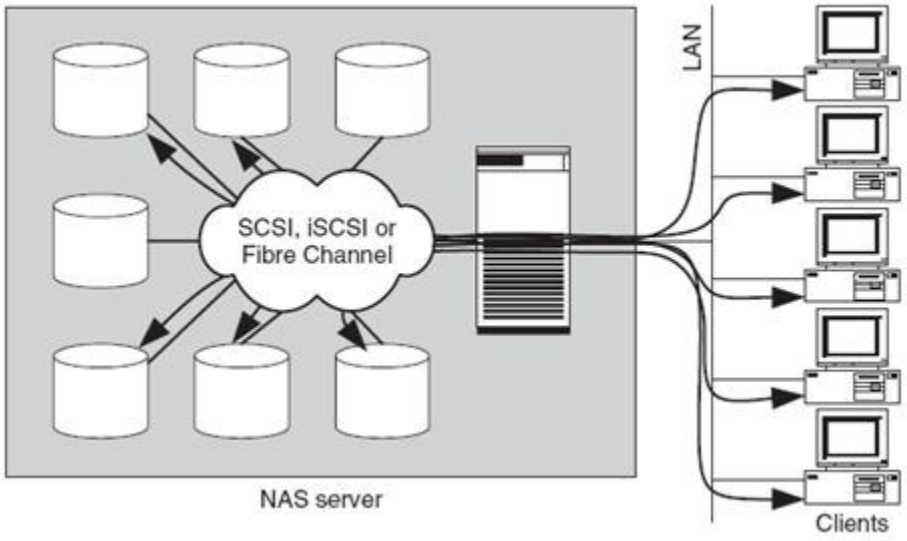
可以在多端口的存储系统上建立 SCSI SAN，如下图所示。



6. NAS

NAS 的全称是 Network-attached storage，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。

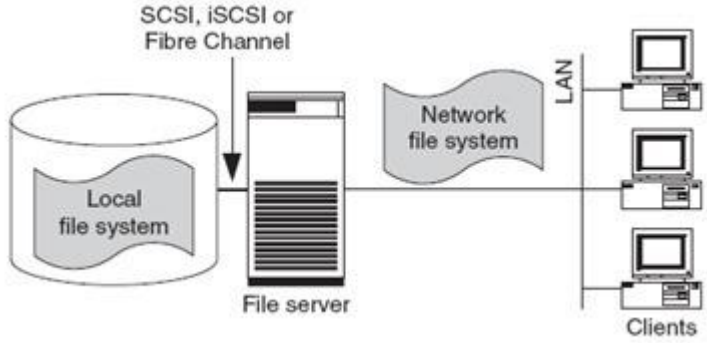
NAS 系统是包含一个或多个硬盘驱动器的网络设备，这些硬盘驱动器通常安排为逻辑的、冗余的存储容器或者 RAID 阵列。



NAS 通常采用 NFS、SMB/CIFS 等网络文件共享协议提供文件存取。

NFS

Network File System（NFS）是一个最早在 1984 年由 Sun 公司提出的网络文件系统协议，它允许客户计算机上的用户按照类似于存取本地文件的方式来存取位于网 络上的文件。类似于其他很多协议，NFS 建立在开放网络计算远程过程调用（OpenNetwork Computing Remote Procedure Call ，ONC RPC）系统之上。NFS 是一个按照 RFCs 定义的公开标准，允许任何人实现。



CIFS/SMB

Common Internet File System（CIFS）又被认为是 ServerMessage Block（SMB），是一个应用程序层的网络协议，这种协议允许主要用于提供对文件、打印机、串口和各种网络节点之间通信的共享存取。它还提供了一种认 证的过程内通信机制。微软使用 CIFS 在所有 Windows 上提供网络功能，Unix/Linux 也通过 SMB 使用 CIFS，Apple 也有一些可以使用 CIFS 的客户端和服务器。因此，它是一个允许各操作系统之间相互协作的协议。

NAS vs DAS

a）DAS 是一个对于已有服务器的简单扩展，并且网络连接不是必需的；NAS 设计为一个容易的并且自给自足的用于在网络上共享文件的解决方案。

b）当都在网络上提供服务时，NAS 比 DAS 的性能更好，因为 NAS 可以为文件服务做精确地调整。

c）当 NAS 和本地（无网络连接的）DAS 进行比较时，NAS 的性能主要取决于网络的速度和拥塞程度。

d）NAS 通常不会按照硬件（CPU、内存、存储）或者软件（扩展、plug-in、附加协议）进行额外的定制，而 DAS 通常会。

e）DAS 和 NAS 都可以通过使用 RAID 或集群来扩展数据的可用性。

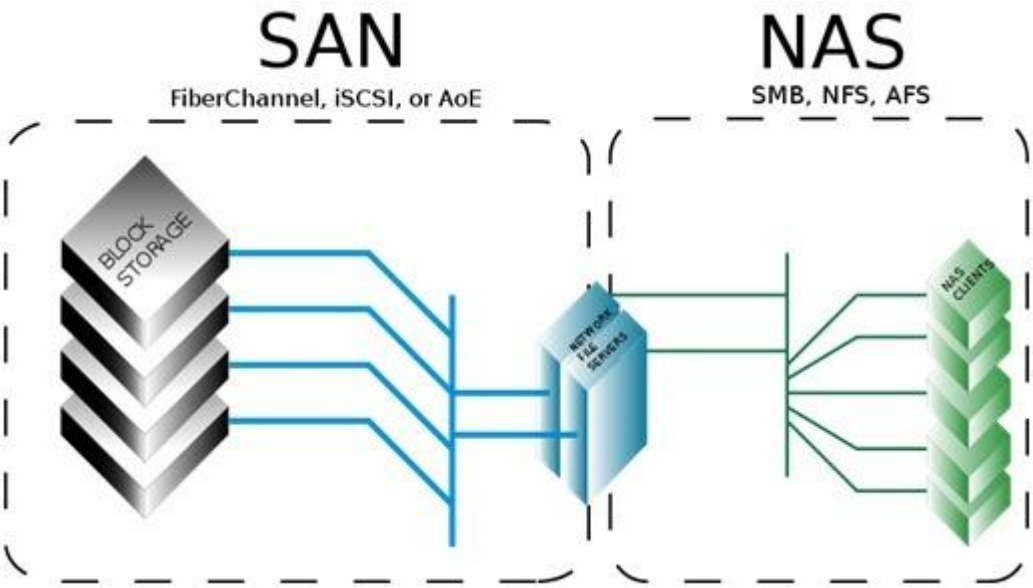
f）NAS 和 DAS 都可以有不同数量的 cache，可以大大的提升性能。

NAS vs SAN

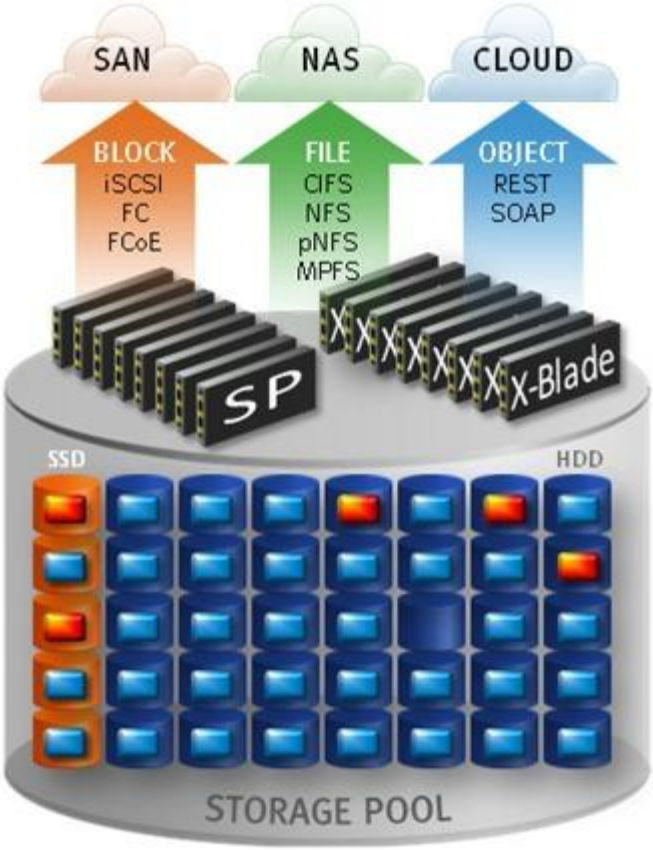
a）NAS 同时提供存储和文件系统；SAN 只提供基于块的存储，而将文件系统留给客户端。

b）NAS 采用的协议是 NFS 和 CIFS/SMB；SAN 采用的协议是 SCSI、Fibre Channel、 iSCSI、ATAover Ethernet（AoE）或者 HyperSCSI。

c）对于 Client OS 来说，NAS 显示为一个文件服务器，客户端可以将网络驱动器映射为该服务器上的共享；而 SAN 对 ClientOS 仍然显示为一个磁盘，该磁盘与客户端的其他本地磁盘一样在磁盘管理工具中可见，并且可以随着文件系统格式化和挂载。

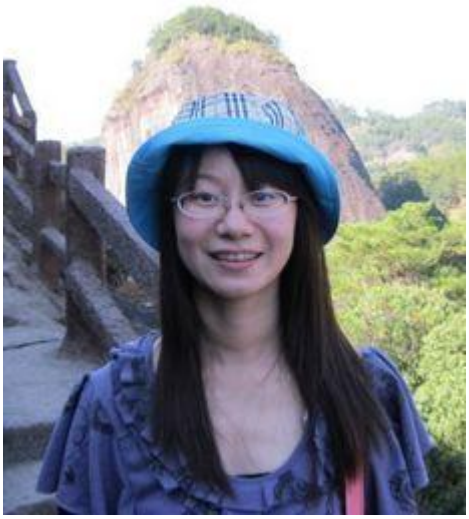


d）但是，NAS 和 SAN 并不是相互排斥的，可以连接起来作为 SAN-NAS 混合体，从同一个系统中同时提供基于文件的协议（NAS）和基于块的协议 （SAN）。如下图所示，EMC 的 Unified Storage 产品 VNX/VNXe 就可以同时提供 NAS 和 SAN 的访问方式。



注： 本文所引有图片多来自 wikipedia 和《Storage Networks Explained》第二版

关于作者



张芸，EMC 中国研究院云平台及应用实验室高级研究员，毕业于西安交通大学，主要关注云计算、存储虚拟化等领域。

微博：http://weibo.com/1756891117

万林涛，EMC 中国研究院云基础架构实验室高级研究员，毕业于南京大学，主要关注于云计算、虚拟化、分布式计算等领域。
微博：http://weibo.com/wanlintao