华为3Com技术有限公司 Huawei-3Com Technologies Co. Ltd.	产品版本 Product version	密级 Confidentiality level
	产品名称 Product name:	Total 96pages 共96页

存储基础知识白皮书

(仅供内部使用）
For internal use only

拟制: Prepared by		日期: Date	yyyy-mm-dd
审核: Reviewed by		日期: Date	yyyy-mm-dd
审核: Reviewed by		日期: Date	yyyy-mm-dd
批准: Granted by		日期: Date	yyyy-mm-dd

华为3Com技术有限公司

Huawei-3Com Technologies Co., Ltd.

修订记录Revision record

日期 Date	修订版本 Revision version	修改描述 change Description	作者 Author
2005-06-28	1.00	初稿完成 initial transmittal	作者名 name
yyyy-mm-dd	1.01	修改 XXX revised xxx	作者名 name
yyyy-mm-dd	1.02	修改 XXX revised xxx	作者名 name
……	……	……	……
yyyy-mm-dd	2.00	修改 XXX revised xxx	作者名 name

目录

第1章 网络存储主要技术

1.1 概述

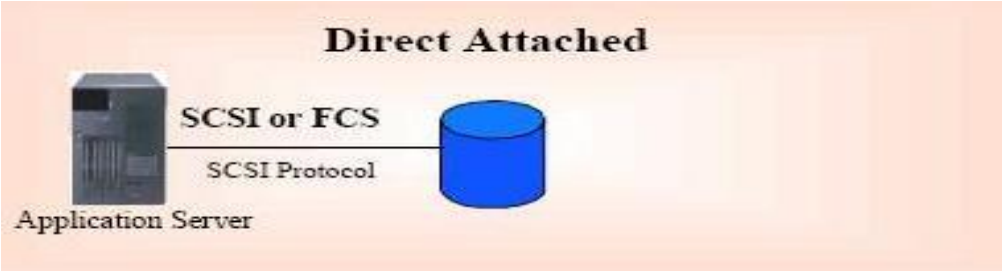
存储系统是整个IT系统的基石，是IT技术赖以存在和发挥效能的基础平台。

早先的存储形式是存储设备（通常是磁盘）与应用服务器其他硬件直接安装于同一个机箱之内，并且该存储设备是给本台应用服务器独占使用的。

随着服务器数量的增多，磁盘数量也在增加，且分散在不同的服务器上，查看每一个磁盘的运行状况都需要到不同的应用服务器上去查看。更换磁盘也需要拆开服务器，中断应用。于是，一种希望将磁盘从服务器中脱离出来，集中到一起管理的需求出现了。不过，一个问题：如何将服务器和盘阵连接起来？

面临这样的问题，有厂商提出了SCSI协议，通过专用的线缆将服务器的总线和存储设备连接起来，通过专门的SCSI指令来实现数据的存储。后来发展到FC协议。这样，多个服务器可以通过SCSI线缆或光纤建立与存储系统的连接。这样的方式，我们称之为直接附加存储（DAS）。

1.2 DAS：直接附加存储



DAS（Direct Attached Storage—直接附加存储）是指将存储设备通过SCSI线缆或光纤通道直接连接到服务器上。

一个SCSI环路或称为SCSI通道可以挂载最多16台设备；

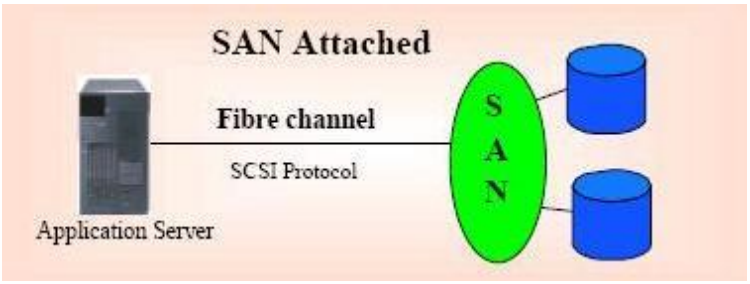
FC可以在仲裁环的方式下支持126个设备；

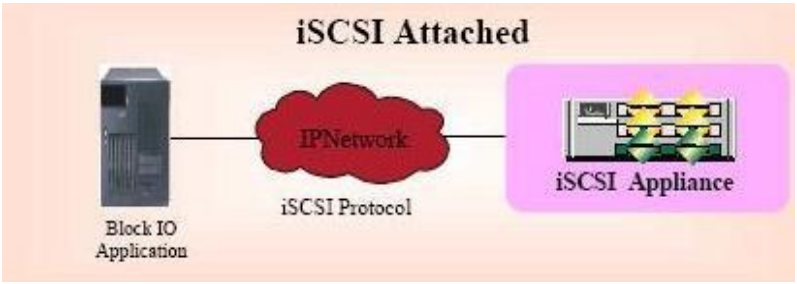
DAS方式实现了机内存储到存储子系统的跨越，但是缺点依然有很多：

- 扩展性差，服务器与存储设备直接连接的方式导致出现新的应用需求时，只能为新增的服务器单独配置存储设备，造成重复投资。
- 资源利用率低，DAS方式的存储长期来看存储空间无法充分利用，存在浪费。不同的应用服务器面对的存储数据量是不一致的，同时业务发展的状况也决定这存储数据量的变化。因此，出现了部分应用对应的存储空间不够用，另一些却有大量的存储空间闲置。
- 可管理性差，DAS方式数据依然是分散的，不同的应用各有一套存储设备。管理分散，无法集中。

异构化严重，DAS方式使得企业在不同阶段采购了不同型号不同厂商的存储设备，设备之间异构化现象严重，导致维护成本据高不下。

1.3 SAN：存储区域网络





1.3.1 什么是 SAN?

SAN（Storage Aera Network ）存储区域网络，是一种通过网络方式连接存储设备和应用服务器的存储构架，这个网络专用于主机和存储设备之间的访问。当有数据的存取需求时，数据可以通过存储区域网络在服务器和后台存储设备之间高速传输。

1.3.2 SAN 的误区

SAN的发展历程较短，从90年代后期兴起，由于当时以太网的带宽有限，而FC协议在当时就可以支持1Gb的带宽，因此早期的SAN存储系统多数由FC存储设备构成，导致很多用户误以为SAN就是光纤通道设备，**其实SAN代表的是一种专用于存储的网络架构，与协议和设备类型无关，随着千兆以太网的普及和万兆以太网的实现，人们对于SAN的理解将更为全面。**

1.3.3 SAN 的组成

SAN由服务器，后端存储系统，SAN连接设备组成；

后端存储系统由SAN控制器和磁盘系统构成，控制器是后端存储系统的关键，它提供存储接入，数据操作及备份，数据共享、数据快照等数据安全管理，及系统管理等一系列功能。

后端存储系统为SAN解决方案提供了存储空间。使用磁盘阵列和RAID策略为数据提供存储空间和安全保护措施。

连接设备包括交换机，HBA卡和各种介质的连接线。

SAN的优点：

- 设备整合，多台服务器可以通过存储网络同时访问后端存储系统，不必为每台服务器单独购买存储设备，降低存储设备异构化程度，减轻维护工作量，降低维护费用；
- 数据集中，不同应用和服务器的数据实现了物理上的集中，空间调整和数据复制等工作可以在一台设备上完成，大大提高了存储资源利用率；
- 高扩展性，存储网络架构使得服务器可以方便的接入现有SAN环境，较好的适应应用变化的需求；

总体拥有成本低，存储设备的整合和数据集中管理，大大降低了重复投资率和长期管理维护成本；

1.3.4 FC SAN 的问题

- 兼容性差，FC协议发展时间短，开发和产品化的大厂商较少，而且厂商之间各自遵循内部标准，导致不同厂商的FC产品之间兼容性和互操作差，即使同一厂商的不同版本不同型号的FC产品也存在类似的问题；
- 成本高昂，FC SAN的成本包括先期设备成本和长期维护成本，由于FC协议在成熟

度和互联性上无法与以太网相比，导致FC协议只能局限于存储系统应用，无法实现大规模推广，这直接导致了FC产品价格的昂贵；同样与FC-SAN相关的所有产品都身价高昂，无论是备份软件的FC-SAN模块，甚至SCSI硬盘简单更换连接口成为FC硬盘，都要翻上几倍的价钱；另外兼容性差也导致了用户无法自己维护FC设备，必须购买昂贵的厂商服务，如果用户的环境中包括多种FC存储设备，用户每年花在FC-SAN的系统保修服务的费用占当年采购成本的15%左右。如果再算上系统安装部署阶段的专业服务费用支出，以5年计算，整个服务费用支出与系统采购达到1:1！

- 扩展能力差，FC-SAN高昂的成本和协议封闭，使得产品的开发、升级、扩容代价高昂。从2000年以来，存储市场中最大的中端部分就一直5年不变地维持着前端两个存储控制器，后端两个（最多四个）光纤环路的结构。不仅产品本身无法进行性能和处理能力扩展，产品型号向上的升级付出的代价几乎相当于购买一套新的设备；
- 异构化严重，各厂商按照自有标准开发各种功能，如快照、复制、镜像等，导致不同厂商存储设备之间功能无法互通，结果又出现的DAS方式的各种问题，重复投资、难以管理的局面

SAN的出现，从根本上是要建立一个开放、高性能、高可靠、高可扩展性的存储资源平台，从而能够应对快速的业务变化和数据增长，然而以上问题使得用户使用网络存储的目标产生了严重的偏离，很多用户甚至开始质疑为什么要放弃DAS而使用昂贵复杂的FC-SAN。

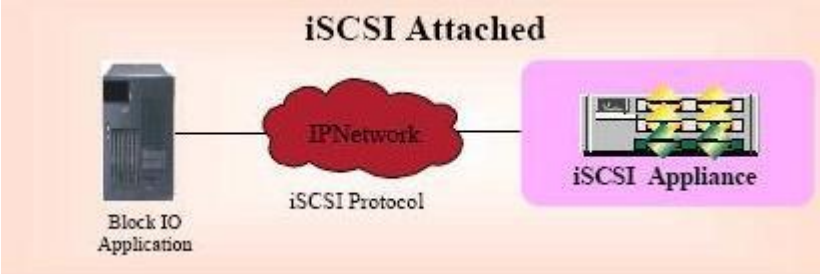
1.3.5 IP SAN

IP网络是一个开放，高性能，高可扩展，可靠性高的网络平台。

- IP网是国际互连网，企业内部网络的主要形式。经过多年发展，IP网络实现了最高的可管理性和互操作性。
- TCP/IP协议弹性强，适应网络的各种变化，无需停止服务即可实网络变更。
- 1G的以太网已经普及，2006年会扩展到10G。FC在2008年才能到4G。
- 不同厂家的IP网设备兼容性好。网络设备采购成本低廉。
- 以太网知识普及，以太网多年的发展培养了无数的网络管理人员。

IP SAN的基本想法是通过高速以太网络连接服务器和后端存储系统。将SCSI指令和数据块经过高速以太网传输，继承以太网的优点，实现建立一个开放、高性能、高可靠性，高可扩展的存储资源平台。

IP SAN



将数据块和SCSI指令通过TCP/IP协议承载，通过千兆/万兆专用的以太网络连接应用服务器和存储设备，这样的解决方案称为IP SAN。

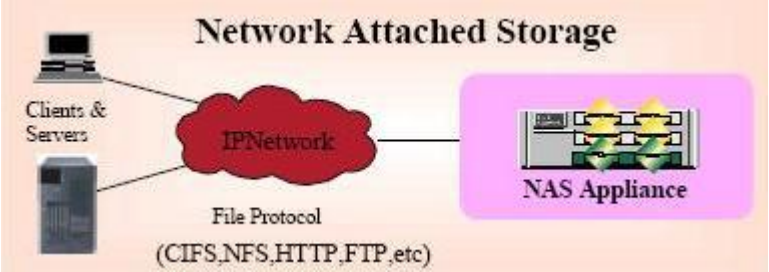
IP SAN遵循IETF的iSCSI标准，通过以太网实现对存储空间的块级访问，由于早先以太网速度，数据安全性以及系统级高容错要求等问题，这一标准经历了三年的认证过程，在包括IBM、HP、SUN、COMPAQ、DELL、Intel、Microsoft、EMC、HDS、Brocade等众多家厂商的努力，和万兆/千兆以太网10GBit Ethernet支撑下，IP SAN/iSCSI已解决了网络瓶颈，数据安全和容错等问题，进入了实用阶段。

IP SAN继承了IP网络的优点：

- 实现弹性扩展的存储网络，能自适应应用的改变。
- 已经验证的传输设备保证运行的可靠性
- 以太网从1G向10G及更高速过渡，只需通过简单的升级便可得到极大的性能提升，并保护投资
- IP跨长距离扩展能力，轻松实现远程数据复制和灾难恢复
- 大量熟悉的网络技术和管理的人才减少培训和人力成本

将以太网的经济性引入存储 降低用户总体拥有成本。

1.4 NAS：网络附加存储



NAS（Network Attached Storage—网络附加存储），是一种文件共享服务。拥有自己的文件系统，通过NFS或CIFS对外提供文件访问服务。

NAS包括存储器件（例如硬盘驱动器阵列、CD或DVD驱动器、磁带驱动器或可移动的存储介质）和专用服务器。专用服务器上装有专门的操作系统，通常是简化的unix/linux操作系统，或者是一个特殊的win2000内核。它为文件系统管理和访问做了专门的优化。专用服务器利用NFS或CIFS，充当远程文件服务器，对外提供文件级的访问。

NAS的优点：

- NAS可以即插即用。
- NAS通过TCP/IP网络连接到应用服务器，因此可以基于已有的企业网络方便连接。
- 专用的操作系统支持不同的文件系统，提供不同操作系统的文件共享。
- 经过优化的文件系统提高了文件的访问效率，也支持相应的网络协议。即使应用服务器不再工作了，仍然可以读出数据。

NAS的缺点：

1、NAS设备与客户机通过企业网进行连接，因此数据备份或存储过程中会占用网络的带宽。这必然会影响企业内部网络上的其他网络应用。共用网络带宽成为限制NAS性能的主要问题。

2、NAS的可扩展性受到设备大小的限制。增加另一台NAS设备非常容易，但是要想将两个NAS设备的存储空间无缝合并并不容易，因为NAS设备通常具有独特的网络标识符，存储空间的扩大上有限。

3、NAS访问需要经过文件系统格式转换，所以是以文件一级来访问。不适和Block级的

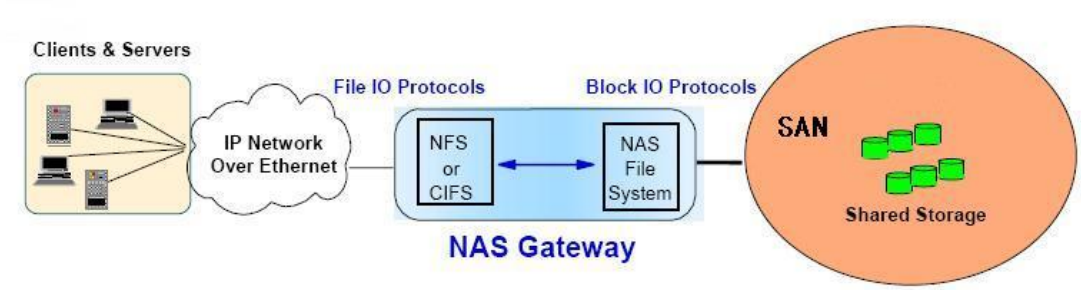
应用，尤其是要求使用裸设备的数据库系统。

1.5 SAN 和 NAS

SAN和NAS经常被视为两种竞争技术，实际上，二者能够很好地相互补充，以提供对不同类型数据的访问。SAN针对海量、面向数据块的数据传输，而NAS则提供文件级的数据访问和共享服务。

尽管这两种技术类似，但严格意义上讲NAS其实只是一种文件服务。

NAS和SAN不仅各有应用场合，也相互结合，许多SAN部署于NAS后台，为NAS设备提供高性能海量存储空间。



NAS和SAN结合中出现了NAS网关这个部件。NAS网关主要由专为提供文件服务而优化的操作系统和相关硬件组成，可以看作是一个专门的文件管理器。NAS网关连接到后端上的SAN上，使的SAN的大容量存储空间可以为NAS所用。因此，NAS网关后面的存储空间可以根据环境的需求扩展到非常大的容量。

“NAS网关”方案主要是在NAS一端增加了可与SAN相连的“接口”，系统对外只有一个用户接口。

NAS网关系统虽然在一定程度上解决了NAS与SAN系统的存储设备级的共享问题，但在文件级的共享问题上却与传统的NAS系统遇到了同样的可扩展性问题。当一个文件系统负载很大时，NAS网关很可能成为系统的瓶颈。

第2章 主要协议和相关技术

关键字：SCSI FC iSCSI

2.1 SCSI

SCSI是小型计算机系统接口（Small Computer System Interface）的简称，于1979首次提出，是为小型机研制的一种接口技术，现在已完全普及到了小型机，高低端服务器以及普通PC上。

SCSI可以划分为SCSI-1、SCSI-2、SCSI-3，最新的为SCSI-3，也是目前应用最广泛的SCSI版本。

1、SCSI-1：1979年提出，支持同步和异步SCSI外围设备；支持7台8位的外围设备，最大数据传输速度为5MB/s。

2、SCSI-2：1992年提出，也称为Fast SCSI，数据传输率提高到20MB/s。

3、SCSI-3：1995年提出，Ultra SCSI（Fast-20）。Ultra 2 SCSI（Fast-40）出现于1997年，最高传输速率可达80MB/s。1998年9月，Ultra 3 SCSI（Utra 160 SCSI）正式发布，最高数据传输率为160MB/s。Ultra 320 SCSI的最高数据传输率已经达到了320MB/s。

2.2 FC（光纤通道）

FC光纤通道：用于计算机设备之间数据传输，传输率达到2G（将来会达到4G）。光纤通道用于服务器共享存储设备的连接，存储控制器和驱动器之间的内部连接。

此图需要更换

协议基本架构：

FC-4 Upper Layer Protocol:SCSI,HIPPI,SBCCS,802.2,ATM,VI,IP

FC-3 common service

FC-2 Framing Protocol /Flow Control

FC-1 Encode/Decode

FC-0 Media:Optical or copper,100MB/sec to 1.062GB/sec

协议层说明：

FC-0：物理层，定制了不同介质，传输距离，信号机制标准，也定义了光纤和铜线接口以及电缆指标

FC-1：定义编码和解码的标准

FC-2：定义了帧、流控制、和服务质量等

FC-3：定义了常用服务，如数据加密和压缩

FC-4：协议映射层，定义了光纤通道和上层应用之间的接口，上层应用比如：串行SCSI 协议，HBA 的驱动提供了FC-4 的接口函数，FC-4 支持多协议，如：FCP-SCSI， FC-IP,FC-VI

协议简介：

FCP-SCSI：

FCP-SCSI:是将SCSI并行接口转化为串行接口方式的协议，应用于存储系统和服务器之

间的数据传输。新的ANSI T10 标准，支持SAN 上存储系统之间通过数据迁移应用来直接移动数据。 FCP-SCSI 提供200MB/s（全双工独占带宽）的传输速率，每连接最远达10 公里，最大16000000 个节点。FCP-SCSI 使用帧传输取代块传输。帧传输以大数据流传输方式传输短的小的事务数据。

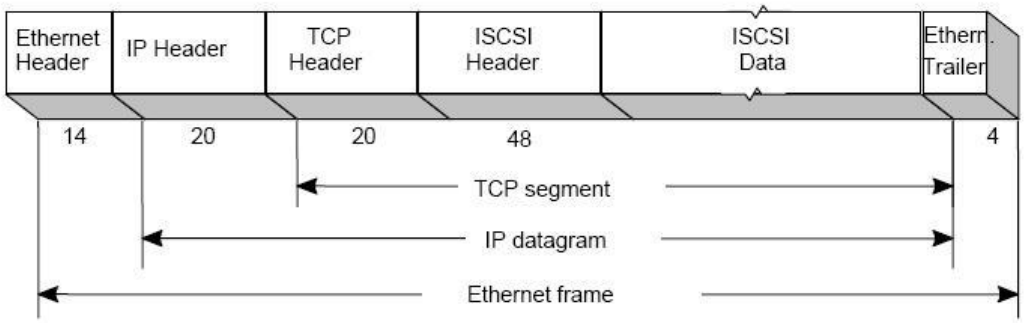
2.3 iSCSI

iSCSI（互联网小型计算机系统接口）是一种在TCP/IP上进行数据块传输的标准。它是由Cisco和IBM两家发起的，并且得到了各大存储厂商的大力支持。iSCSI可以实现在IP网络上运行SCSI协议，使其能够在诸如高速千兆以太网上进行快速的数据存取备份操作。

iSCSI标准在2003年2月11日由IETF（互联网工程任务组）认证通过。iSCSI继承了两大最传统技术：SCSI和TCP/IP协议。这为iSCSI的发展奠定了坚实的基础。

基于iSCSI的存储系统只需要不多的投资便可实现SAN存储功能，甚至直接利用现有的TCP/IP网络。相对于以往的网络存储技术，它解决了开放性、容量、传输速度、兼容性、安全性等问题，其优越的性能使其备受始关注与青睐。

iSCSI的数据包结构：



工作流程：

iSCSI系统由SCSI适配器发送一个SCSI命令。

命令封装到TCP/IP包中并送入到以太网络。

接收方从TCP/IP包中抽取SCSI命令并执行相关操作。

把返回的SCSI命令和数据封装到TCP/IP包中，将它们发回到发送方。

系统提取出数据或命令，并把它们传回SCSI子系统。

安全性描述：

iSCSI协议本身提供了QoS及安全特性。

可以限制initiator仅向target列表中的目标发登录请求，再由target确认并返回响应，之后才允许通信；

通过IPSec将数据包加密之后传输，包括数据完整性、确定性及机密性检测等；

iSCSI的优势

（1）广泛分布的以太网为iSCSI的部署提供了基础。

（2）千兆/万兆以太网的普及为iSCSI提供了更大的运行带宽。

（3）以太网知识的普及为基于iSCSI技术的存储技术提供了大量的管理人才。

（4）由于基于TCP/IP网络，完全解决数据远程复制（Data Replication）及灾难恢复（Disaster Recover）等传输距离上的难题。

（5）得益于以太网设 备的价格优势和TCP/IP网络的开放性和便利的管理性，设备扩充和应用调整的成本付出小。

2.4 iSCSI 与光纤通道的比较

从传输层看，光纤通道的传输采用其FC协议，iSCSI采用TCP/IP协议。

FC协议与现有的以太网是完全异构的，两者不能相互接驳。因此光纤通道是具有封闭性的，而且不仅与现有的企业内部网络（以太网）接入，也与其他不同厂商的光纤通道网络接入（由于厂家对FC标准的理解的异样，FC设备的兼容性是一个巨大的难题）。因此，对于以后存储网络的扩展由于兼容性的问题而成为了难题。而且，FC协议由于其协议特性，网络建完后，加入新的存储子网时，必须要重新配置整个网络，这也是FC网络扩展的障碍。

iSCSI基于的TCP/IP协议，它本身就运行于以太网之上，因此可以和现有的企业内部以太网无缝结合。TCP/IP网络设备之间的兼容性已经无需讨论，迅猛发展的internent网上运行着全球无数家网络设备厂商提供的网络设备，这是一个最好的佐证。

从网络管理的角度看，运行FC协议的光网络，其技术难度相当之大。其管理采用了专

有的软件，因此需要专门的管理人员，且其培训费用高昂。TCP/IP网络的知识通过这些年的普及，已有大量的网络管理人才，并且，由于支持TCP/IP的设备对协议的支持一致性好，即使是不同厂家的设备，其网络管理方法也是基本一致的。

FC运行于光网络之上，其速度是非常快的，现在已经达到了2G的带宽，这也是它的主要优势所在。下一代的FC标准正在制定当中，其速度可以达到4G，

今天的千兆以太网已经在普及当中，这也是基于TCP/IP的iSCSI协议进入实用的保证。得益于优秀的设计，以太网从诞生到现在，遍及了所有有网络的地方，到现在依然表现出非凡的生命力，在全球无数网络厂商的共同努力下，以太网的速度稳步提升，千兆网络已经实际应用，万兆网络呼之欲出，以太网的主要部件交换机路由器均已有万兆级别的产品。随着产品的不断丰富，以及设备厂商间的剧烈竞争，其建设成本在不断下降，万兆网络的普及已日益临近。当iSCSI以10Gb的高速传输数据时，基于iSCSI协议的存储技术将无可争议的成为网络存储的王者。

第3章 文件系统相关知识

3.1 什么是文件系统

文件系统定义了把文件存储于磁盘时所必须的数据结构及磁盘数据的管理方式。我们知道，磁盘是由很多个扇区（Sector）组成的，如果扇区之间不建立任何的关系，写入其中的文件就无法访问，因为无法知道文件从哪个扇区开始，文件占多少个扇区，文件有什么属性。为了访问磁盘中的数据，就必需在扇区之间建立联系，也就是需要一种逻辑上的数据存储结构。建立这种逻辑结构就是文件系统要做的事情，在磁盘上建立文件系统的过程通常称为“格式化”。

以Windows平台下最常见的FAT文件系统为例。FAT文件系统有两个重要的组成部分：FAT表（File Allocation Table）和数据存储区。FAT表是FAT文件系统的名称来源，它定义;了存储数据的簇（Cluster，由2的n次方个Sector组成，n值根据分区大小而定，需综合考虑数据存取效率和存储空间的利用率）之间的链接关系，这种链接关系是一个单向链表，指向0xFF表示结束。依据一个簇编号所用bit数的不同，可分为FAT12、FAT16和FAT32文件系统。数据区存储的数据包含文件目录项（Directory Entries）和文件数据。文件目录项存储的是一个文件或目录的属性信息，包括文件名称（把目录也看成是文件）、读写属性、文件大小、创建时间、起始簇编号等，一个目录下的每个子目录和文件都对应一个表项记录。文件目录项以固定32字节的长度存储，以树型结构管理，其中根目录的位置是确定的。也就是说，根据分区根目录可以找到下级子目录和文件的起始簇编号，根据下级子目录又可以找到更下级目录或文件的起始簇编号。可见，FAT表和文件目录项是为了文件的访问和管理而建立的。应用程序要访问一个文件时，根据文件路径（逻辑分区号＋目录，如F:\software）和文件名称（如setup.exe）可从文件目录项中获得存储文件数据的起始簇号，之后从FAT表查询这个簇号对应的链表，就可以获得该文件对应的全部簇编号。从这些簇中读出全部数据，就得到一个完整的文件。

一般来说，文件系统是和操作系统紧密结合在一起的，不同的操作系统使用不同的文件系统，但有时为了兼容，不同操作系统也使用相同的文件系统。

3.2 主流文件系统和特点

在Windows系列操作系统中，MS-DOS和Windows 3.x使用FAT16文件系统，默认情况下Windows 98也使用FAT16，Windows 98和Windows Me可以同时支持FAT16、FAT32两种文件系统，Windows NT则支持FAT16、NTFS两种文件系统，Windows 2000可以支持FAT16、FAT32、NTFS三种文件系统.每一种文件系统提供的功能与特点各不相同。比如FAT32文件系统。，采用32位的文件分配表，磁盘的管理能力大为增强。但由于文件分配表的增大，性能相对来说有所下降。此外，这个版本的文件系统不能向下兼容。

NTFS是随着 Windows NT操作系统而产生的，它的优点和FAT文件系统相比是有更好的安全性和稳定性，在使用中不易产生文件碎片，NTFS分区对用户权限作出了非常严格的限制，同时它还提供了容错结构日志，从而保护了系统的安全。但NTFS分区格式的兼容性不好，Windows 98/ME操作系统均不能直接访问该分区。

对于超过4GB以上的硬盘，使用NTFS分区，可以减少磁盘碎片的数量，大大提高硬盘的利用率；NTFS可以支持的文件大小可以达到64GB，远远大于FAT32下的4GB；支持长文件名，支持的最大分区为 2TB。

在Linux系统中，每个分区都是一个文件系统，都有自己的目录层次结构。Linux的最重要特征之一就是支持多种文件系统,并可以和许多其它种操作系统共存。

随着Linux的不断发展，它所支持的文件格式系统也在迅速扩充。特别是Linux 2.4内核正式推出后，出现了大量新的文件系统. Linux系统可以支持十多种文件系统类型包括：JFS、ext、ext2、ext3、ISO9660、XFS、Minx、MSDOS、UMSDOS、VFAT、NTFS、HPFS、NFS、SMB、SysV、PROC等。

各主流操作系统和平台的文件系统名称和特点如下表所示

操作系统	文件系统	特点
Windows 95、Windows 98、OSR2、Windows 98 SE、Windows Me、Windows 2000和Windows XP	Fat文件系统 FAT12/FAT16和FAT32	可以允许多种操作系统访问，如MS-DOS、Windows 3.x、Windows 9x、Windows NT和OS/2等。这一文件系统在使用时遵循8.3命名规则(即文件名最多为8个字符，扩展名为3个字符)。 最大的限制在于兼容性方面，Fat32不能保持向下兼容。 当分区小于512M时，Fat32不会发生作用。 单个文件不能大于4G。
Windows NT/2000	NTFS文件系统	支持文件系统故障恢复，尤其是大存储媒体、长文件名。分区大小可以达到2TB。通过使用标准的事物处理日志和恢复技术来保证分区的一致性。 只能被Windows NT/2000所识别，不能被FAT文件系统所存取
Windows longhorn	Winfs	用以组织、搜索和共享多种多样的信息的存储平台。WinFS被设计为在无结构文件和数据库数据之间建立起更好的互操作性，从而提供快捷的文件浏览和搜索功能
Linux	Ext2/ ext3/ XFS等文件系统	是一种日志式文件系统。日志式文件系统的优越性在于：由于文件系统都有快取层参与运作，如不使用时必须将文件系统卸下，以便将快取层的资料写回磁盘中。因此每当系统要关机时，必须将其所有的文件系统全部卸下后才能进行关机
UNIX 系统	NFS	网络文件系统，允许多台计算机之间共享文件系统，易于从所有这些计算机存放文件
Windows 系列	CIFS	网络文件系统，允许多台计算机之间共享文件系统，易于从所有这些计算机存放文件
AIX	JFS	具有可伸缩性和健壮性，与非日志文件系统相比，它的优点是其快速重启能力：Jfs 能够在几秒或几分钟内就把文件系统恢复到一致状态。为满足服务器（从单处理器系统到高级多处理器和群集系统）的高吞吐量和可靠性需求而设计的。使用数据库日志处理技术，jsf 能在几秒或几分钟之内把文件系统恢复到一致状态。
SCO UnixWare，	Vxfs UFS	日志式文件系统.建立文件的索引区，将操作记录在事件日志中，当系统发生意外时，能让系统迅速、完
	Vxfs	

3.3 NFS 和 CIFS 网络文件系统工作原理和特点

NFS (Network File System，网络文件系统)是当前主流异构平台共享文件系统之一.主要应用在UNIX环境下。 最早是由SUN microsystem开发，现在能够支持在不同类型的系统之间通过网络进行文件共享，广泛应用在FreeBSD、SCO、Solaris等等异构操作系统平台， 允许一个系统在网络上与它人共享目录和文件。通过使用NFS，用户和程序可以象访问本地文件一样访问远端系统上的文件,使得每个计算机的节点能够像使用本地资源一样方便地使用网上资源。换言之，NFS 可用于不同类型计算机、操作系统、网络架构和传输协议运行环境中的网络文件远程访问和共享。

NFS的工作原理是使用客户端/服务器架构，由一个客户端程序和服务器程序组成。服务器程序向其它计算机提供对文件系统的访问，其过程就叫做“输出”。NFS 客户端程序对共享文件系统进行访问时，把它们从 NFS 服务器中“输送”出来。文件通常以“块” 为单位进行传输. 其尺寸是 8K (虽然它可能会将操作分成更小尺寸的分片).NFS 传输协议用于服务器和客户机之间文件访问和共享的通信,从而使客户机远程地访问保存在存储设备上的数据。

CIFS（Common Internet File Syste，公共互联网文件系统）是当前主流异构平台共享文件系统之一。主要应用在NT/Windows环境下，是由Microsoft公司开发。其工作原理是让CIFS协议运行于TCP/IP通信协议之上，让Unix计算机可以在网络邻居上被Windows计算机看到。

共享文件系统特点：

- 异构平台下的文件共享:不同平台下的多个客户端可以很容易的共享 NAS 中的同一个文件。
- 充分利用现有的 LAN 网络结构，保护现有投资。
- 容易安装，使用和管理都很方便，实现即插即用。
- 广泛的连接性：由于基于 IP/Ethernet 以及标准的 NFS 和 CIFS，可以适应复杂的网络环境。
- 内部资源的整合：可以将内部的磁盘整合成一个统一的存储池，以卷的方式提供给不同的用户，每一个卷可以格式化成不同的文件系统

- 允许应用进程打开一个远地文件，并能够在该文件的某一个特定的位置上开始读写数据。NFS 可使用户只复制一个大文件中的一个很小的片段，而不需复制整个大文件，在网络上传送的只是少量的修改数据。

需要注意的是，CIFS和NFS虽然同样也是文件系统（File System），但它并不能用于在磁盘中存储和管理数据，它定义的是通过TCP/IP网络传输文件时的文件组织格式和数据传输方式。利用CIFS和NFS共享文件实际涉及到两次的文件系统转换。客户端从服务器端申请一个文件时，服务器端首先从本地读出文件（本地文件系统格式），并以NFS/CIFS的格式封装成IP报文并发送给客户端。客户端收到IP报文以后，把文件存储与本地磁盘中（本地文件系统格式）。

3.4 存储系统与文件系统的关系

提到NAS，通常会想到传统的NAS设备，它具有自己的文件系统，具有较大的存储容量，具有一定的文件管理和服务功能。NAS设备和客户端之间通过IP网络连接，基于NFS/CIFS协议在不同平台之间共享文件，数据的传输以文件为组织单位。

虽然NAS设备常被认为是一种存储架构，但NAS设备最核心的东西实际上在存储之外，那就是文件管理服务。从功能上来看，传统NAS设备就是一个带有DAS存储的文件服务器。从数据的IO路径来看，它的数据IO发生在NAS设备内部，这种架构与DAS毫无分别。而事实上，很多NAS设备内部的文件服务模块与磁盘之间是通过SCSI总线连接的。至于通过NFS/CIFS共享文件，完全属于高层协议通信，根本就不在数据IO路径上，所以数据的传输不可能以块来组织。正是由于这种功能上的重叠，在SAN出现以后，NAS头设备（或NAS网关）逐渐发展起来，NAS over SAN的方案越来越多，NAS回归了其文件服务的本质。

由此可知，NAS与一般的应用主机在网络层次上的位置是相同的，为了在磁盘中存储数据，就必须要建立文件系统。有的NAS设备采用专有文件系统，而有的NAS设备则直接借用其操作系统支持的文件系统。由于不同的OS平台之间文件系统不兼容，所以NAS设备和客户端之间就采用通用的NFS/CIFS来共享文件。

至于SAN，它提供给应用主机的就是一块未建立文件系统的“虚拟磁盘”。在上面建立什么样的文件系统，完全由主机操作系统确定。

第4章 RAID 技术

4.1 RAID 概述

RAID为廉价磁盘冗余阵列（Redundant Array of Inexpensive Disks），RAID技术将一个个单独的磁盘以不同的组合方式形成一个逻辑硬盘，从而提高了磁盘读取的性能和数据的安全性。不同的组合方式用RAID级别来标识。

RAID技术是由美国加州大学伯克利分校D.A. Patterson教授在1988年提出的，作为高性能、高可靠的存储技术，在今天已经得到了广泛的应用。

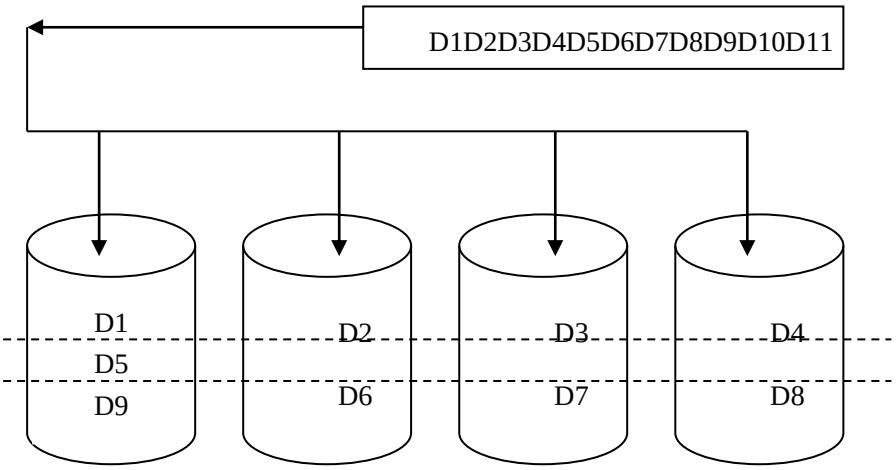
4.2 RAID 级别

RAID技术经过不断的发展，现在已拥有了从 RAID 0 到 5等6种明确标准级别的RAID级别。另外，其他还有6、7、10（RAID 1与RAID 0的组合）、01（RAID 0与RAID 1的组合）、30（RAID 3与RAID 0的组合）、50（RAID 0与RAID 5的组合）等。

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，下面将介绍如下RAID级别：0、1、2、3、4、5、6、01、10。

4.2.1 RAID0

RAID0也称为条带化（stripe），将数据分成一定的大小顺序的写到阵列的磁盘里，RAID0可以并行的执行读写操作，可以充分利用总线的带宽，理论上讲，一个由N个磁盘组成的RAID0系统，它的读写性能将是单个磁盘读取性能的N倍。且磁盘空间的存储效率最大（100％）RAID0有一个明显的缺点：不提供数据冗余保护，一旦数据损坏，将无法恢复。

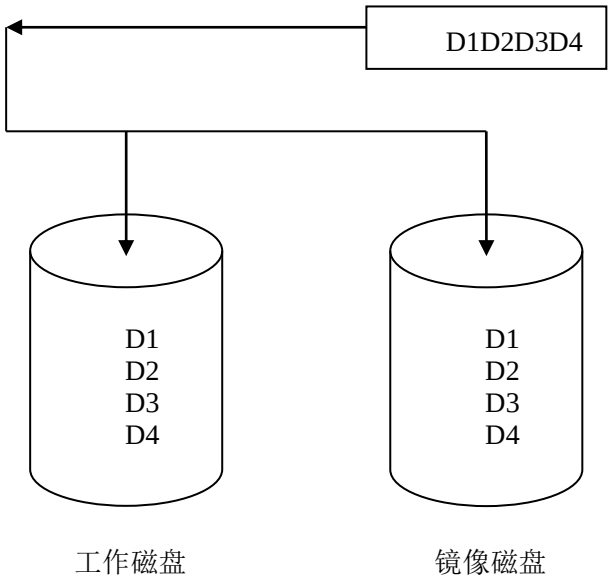


如图所示：系统向RAID0系统（四个磁盘组成）发出的I/O数据请求被转化为4项操作，其中的每一项操作都对应于一块物理硬盘。通过建立RAID 0，原先顺序的数据请求被分散到四块硬盘中同时执行。从理论上讲，四块硬盘的并行操作使同一时间内磁盘读写速度提升了4倍。 但由于总线带宽等多种因素的影响，实际的提升速率会低于理论值，但是，大量数据并行传输与串行传输比较，性能必然大幅提高。

RAID0应用于对读取性能要求较高但所存储的数据为非重要数据的情况下。

4.2.2 RAID1

RAID1成为镜像（mirror），它将数据完全一致的分别写到工作磁盘和镜像磁盘，因此它的磁盘空间利用率为50%，在数据写入时时间会有影响，但是读的时候没有任何影响，RAID0提供了最佳的数据保护，一旦工作磁盘发生故障，系统自动从镜像磁盘读取数据，不会影响用户工作。



RAID1应用于对数据保护极为重视的应用。

4.2.3 RAID2

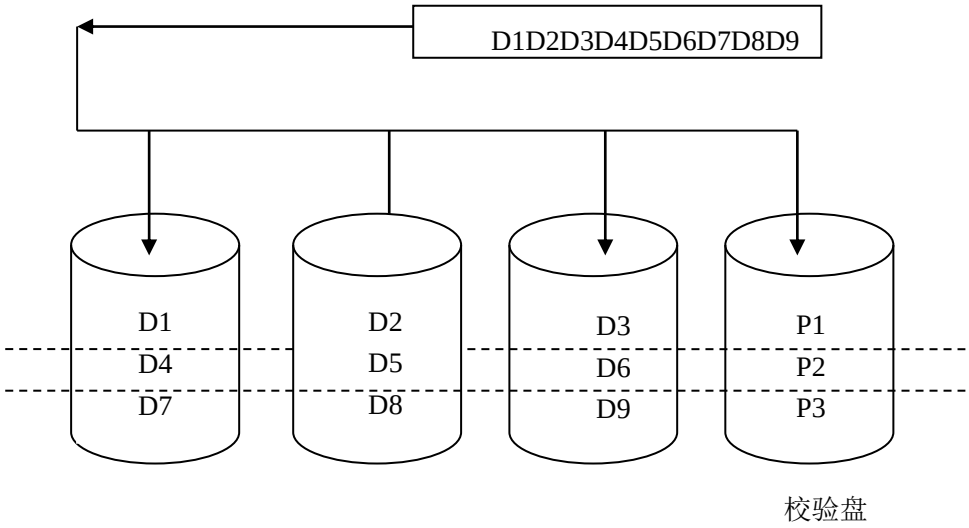
RAID2称为纠错海明码磁盘阵列，阵列中序号为2N的磁盘（第1、2、4、6……）作为校验盘，其余的磁盘用于存放数据，磁盘数目越多，校验盘所占比率越少。RAID2在大数据存储额情况下性能很高，RAID2的实际应用很少。

4.2.4 RAID3

RAID3采用一个硬盘作为校验盘，其余磁盘作为数据盘，数据按位或字节的方式交叉的存取到各个数据盘中。不同磁盘上同一带区的数据做异或校验，并把校验值写入到校验盘中。RAID3系统在完整的情况下读取时没有任何性能上的影响，读性能与RAID0一致，却提供了数据容错能力，但是，在写时性能大为下降，因为每一次写操作，即使是改动某个数据盘上的一个数据块，也必须根据所有同一带区的数据来重新计算校验值写入到校验盘中，一个写操作包含了写入数据块，读取同一带区的数据块，计算校验值，写入校验值等操作，系统开销大为增加。

当RAID3中有数据盘出现损坏，不会影响用户读取数据，如果读取的数据块正好在损坏

的磁盘上，则系统需要读取所有同一带区的数据块，然后根据校验值重新构建数据，系统性能受到影响。



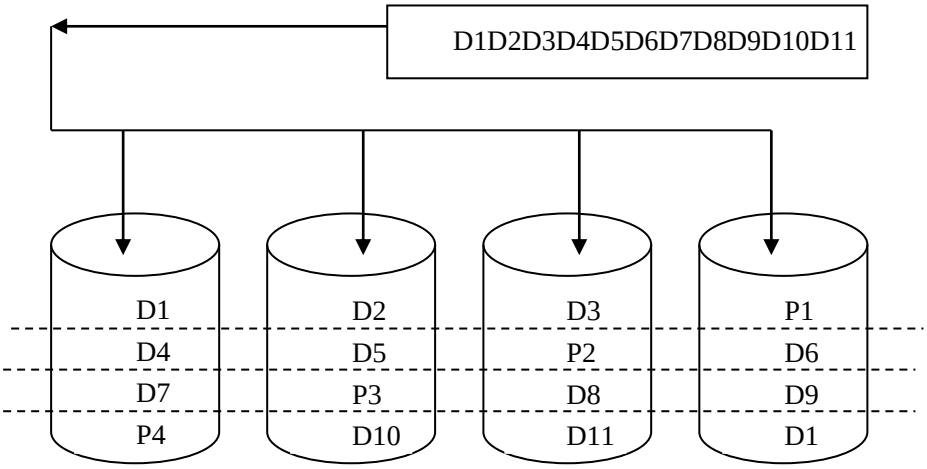
RAID3的校验盘在系统接受大量的写操作时容易形成性能瓶颈，因而适用于有大量读操作如web系统以及信息查询等应用或持续大块数据流（例如非线性编辑）的应用。

4.2.5 RAID4

RAID4与RAID3基本一致，区别在于条带化的方式不一样，RAID4按照块的方式存放数据，所以在写操作时只涉及两块磁盘，数据盘和校验盘，提高了系统的IO性能。但面对随机的分散的写操作，单一的校验盘往往成为性能瓶颈。

4.2.6 RAID5

RAID5与RAID3的机制相似，但是数据校验的信息被均匀的分散到的阵列的各个磁盘上，这样就不存在并发写操作时的校验盘性能瓶颈。阵列的磁盘上既有数据，也有数据校验信息，数据块和对应的校验信息会存储于不同的磁盘上，当一个数据盘损坏时，系统可以根据同一带区的其他数据块和对应的校验信息来重构损坏的数据。



RAID 5可以理解为是RAID 0和RAID 1的折衷方案。RAID 5可以为系统提供数据安全保障，但保障程度要比RAID1低而磁盘空间利用率要比RAID1高。RAID 5具有和RAID 0相近似的数据读取速度，只是多了一个奇偶校验信息，写入数据的速度比对单个磁盘进行写入操作稍慢。同时由于多个数据对应一个奇偶校验信息，RAID 5的磁盘空间利用率要比RAID 1高，存储成本相对较低。

RAID5在数据盘损坏时的情况和RAID3相似，由于需要重构数据，性能会受到影响。

4.2.7 RAID6

RAID 6提供两级冗余，即阵列中的两个驱动器失败时，阵列仍然能够继续工作。

一般而言，RAID 6的实现代价最高，因为RAID 6不仅要支持数据的恢复，又要支持校验的恢复，这使RAID 6控制器比其他级R A I D更复杂和更昂贵。

1. RAID 6的校验数据

当对每个数据块执行写操作时， RAID 6做两个独立的校验计算，因此，它能够支持两个磁盘的失败。为了实现这个思想，目前基本上有两个已经接受的方法：

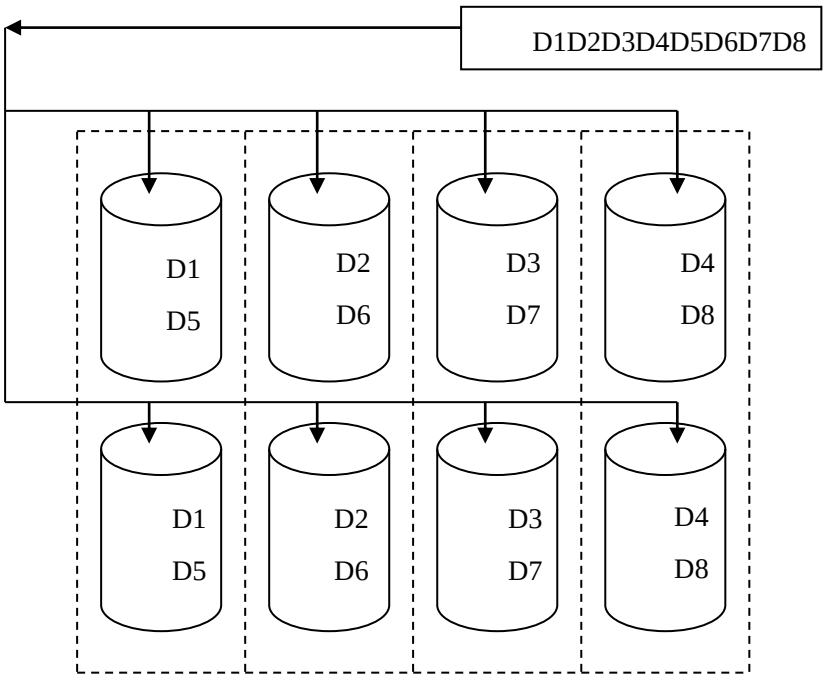
- 使用多种算法，如X O R和某种其他的函数。
- 在不同的数据分条或者磁盘上，使用排列的数据。

2. RAID 6的一维冗余

RAID 6的第一种方法是用两种不同的方法计算校验数据。实现这个思想最容易的方法之一是用两个校验磁盘支持数据磁盘，第一个校验磁盘支持一种校验算法，而第二个磁盘支持另一种校验算法，使用两种算法称为P + Q校验。一维冗余是指使用另一个校验磁盘，但所包含的分块数据是相同的。例如，P校验值可能由X O R函数产生，这样，Q校验函数需要是其他的某种操作，一个很有力的侯选者是Reed Solomon误差修正编码的变体，这个误差修正编码一般用于磁盘和磁带驱动器。假如两个磁盘失败，那么，通过求解带有两个变量的方程，可以恢复两个磁盘上的数据，这是一个代数方法，可以由硬件辅助处理器加速求解。

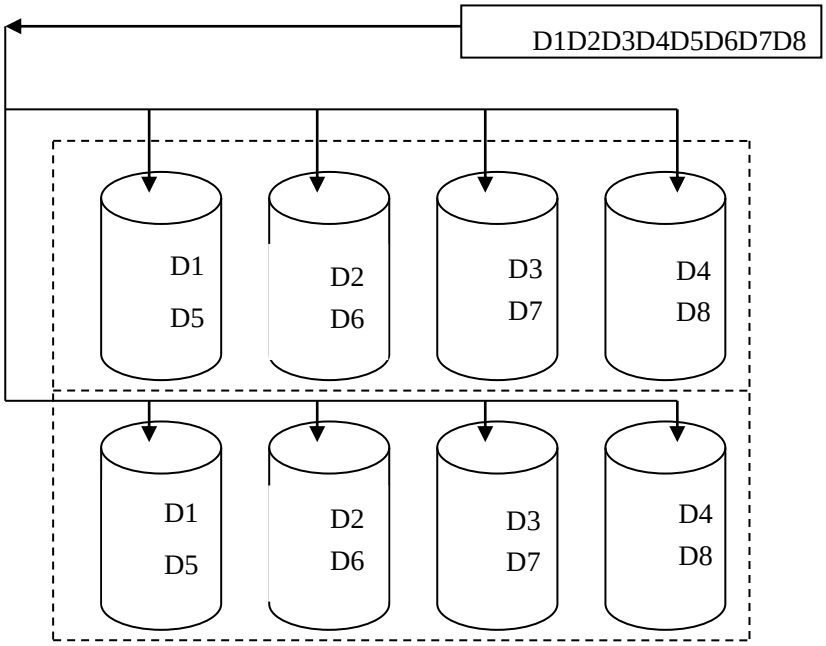
4.2.8 RAID10

RAID10是RAID1和RAID0的结合，也称为RAID（0+1），先做镜像然后做条带化，既提高了系统的读写性能，有提供了数据冗余保护，RAID10的磁盘空间利用率和RAID1是一样的，为50%。RAID10适用于既有大量的数据需要存储，有对数据安全性有严格要求的领域，比如金融，证券等。



4.2.9 RAID01

RAID01也是RAID0和RAID1的结合，但它是对条带化后的数据进行镜像。但与RAID10不同，一个磁盘的丢失等同于整个镜像条带的丢失，所以一旦镜像盘失败，则存储系统成为一个RAID-0 系统（即只有条带化）。RAID01的实际应用非常少。



4.2.10 JBOD

JBOD(Just Bundle Of Disks)译成中文可以是"简单磁盘捆绑"，通常又称为Span。 JBOD不是标准的RAID级别，它只是在近几年才被一些厂家提出，并被广泛采用。

Span是在逻辑上把几个物理磁盘一个接一个串联到一起，从而提供一个大的逻辑磁盘。Span上的数据简单的从第一个磁盘开始存储， 当第一个磁盘的存储空间用完后， 再依次从后面的磁盘开始存储数据。

Span存取性能完全等同于对单一磁盘的存取操作。Span也不提供数据安全保障。它只是简单的提供一种利用磁盘空间的方法，Span的存储容量等于组成Span的所有磁盘的容量的总和。

4.3 不同 RAID 级别对比

在各个raid级别中，使用最广泛的是raid0，raid1，raid10，raid5

RAID-0，将数据分成条带顺序写入一组磁盘中。RAID-0 不提供冗余功能，但是它却提供了卓越的吞吐性能，因为读写数据是在一组磁盘中的每个磁盘上同时处理的，吞吐性能远远超过单个磁盘的读写。

RAID-1，每次写操作都将分别写两份到数据盘和校验盘上，每对数据盘和校验盘成为镜像磁盘组。也可使用并发的方式来读数据时，提高吞吐性能。如果镜像磁盘组中某个磁盘出错，则数据可以从另外一块磁盘获得，而不会影响系统的性能，然后，使用一块备用磁盘将健康磁盘中的数据复制出来然后这两块磁盘又组成新的镜像组。

RAID1/0，即RAID1 与RAID0 的结合，既做镜像又做条带化，数据先镜像再做条带化。这样数据存储既保证了可靠性，又极大地提高了吞吐性能。

RAID-0/1 也是RAID0 与RAID1 的结合，但它是对条带化后的数据进行镜像。但与RAID10 不同，一个磁盘的丢失等同于整个镜像条带的丢失，所以一旦镜像盘失败，则存储系统成为一个RAID-0 系统（即只有条带化）。

RAID-5 是将数据校验循环分散到各个磁盘中，它像RAID-0 一样将数据条带化分散写到一组磁盘中，但同时它生成校验数据做为冗余和容错使用。校验磁盘包含了所有条带的数据的校验信息。RAID-5 将校验信息轮流地写入条带磁盘组的各个磁盘中，即每个磁盘上既有数据信息又同时有校验信息，RAID-5 的性能得益于数据的条带化，但是某个磁盘的失败却将引起整个系统的下降，这是因为系统将在承担读写任务的同时，重新构建和计算出失败磁盘上的数据，此时要使用备用磁盘对失败磁盘的数据重建恢复整个系统的健康。

从一个普通应用来讲，要求存储系统具有良好的IO性能同时也要求对数据安全做好保护工作，所以raid10和raid5应该成为我们重点关注的对象。

下面从IO性能，数据重构及对系统性能的影响，数据安全保护等方面，结合磁盘现状来分析两种技术的差异。

IO的性能：读操作上raid10和raid5是相当的，RAID-5 在一些很小数据的写操作（如比每个条带还小的小数据）需要2 个读、2 个写，还有2 个XOR 操作，对于单个用户的写操

作，在新数据应用之前必须将老的数据从校验盘中移除，整个的执行过程是这样：读出旧数据，旧数据与新数据做XOR，并创建一个即时的值，读出旧数据的校验信息，将即时值与校验数据进行XOR，最后写下新的校验信息。为了减少对系统的影响，大多数的RAID5 都读出并将整个条带（包括校验条带）写入缓存，执行2 个XOR 操作，然后发出并行写操作（通常对整个条带），即便了进行了上述优化，系统仍然需要为这种写操作进行额外的读和XOR操作。小量写操作困难使得RAID-5 技术很少应用于密集写操作的场合，如回滚字段及重做日志。当然，也可以将存储系统的条带大小定义为经常读写动作的数据大小，使之匹配，但这样会限制系统的灵活性，也不适用于企业中其它的应用。

对于raid10，由于不存在数据校验，每次写操作只是单纯的执行写操作。应此在写性能上raid10要好于raid5。

数据重构：

对于raid10，当一块磁盘失效时，进行数据重构的操作只是复制一个新磁盘，如果假定磁盘的容量为250G，那么复制的数据量为250G。

对于raid5的存储阵列，则需要从每块磁盘中读取数据，经过重新计算得到一块硬盘的数据量，如果raid5是以4+1的方式组建，每块磁盘的容量也为250G，那么，需要在剩余的4个磁盘中读出总共是1000G的数据量计算得出250G的数据。

从这点来看，raid5在数据重构上的工作负荷和花费的时间应该远大于raid10，负荷变大将影响重构期间的性能，时间长意味再次出现数据损坏的可能性变大。

数据安全保护：

raid10系统在已有一块磁盘失效的情况下，只有出现该失效盘的对应镜像盘也失效，才会导致数据丢失。其他的磁盘失效不会出现数据丢失情况。

Raid5系统在已有一块磁盘失效的情况下，只要再出现任意的一块磁盘失效，都将导致数据丢失。

从综合来看，raid10和raid5系统在出现一块磁盘失效后，进行数据重构时，raid5需耗费的时间要比raid10长，同时重构期间系统负荷上raid5要比raid10高，同时raid5出现数据丢失的可能性要比raid10高，因此，数据重构期间，raid5系统的可靠性远比raid10来的低。

Raid5在磁盘空间率用率上比raid10高，raid5的空间利用率是（N-1）/ N （N为阵列的磁盘数目），而raid10的磁盘空间利用率仅为50％。

但是结合磁盘来考虑，今天的硬盘厂商所生产的ATA或SATA硬盘的质量已经可以承担企业级的应用，并且，容量的增加幅度相当大，目前已经可以实现单个磁盘400G的存储容量。SCSI硬盘由于要求高转速而使用小直径盘片，容量的增加相对缓慢。ATA磁盘相对SCSI磁盘拥有成本也要小很多。

应此，在采用价格昂贵的FC或SCSI硬盘的存储系统中，对于预算有限同时数据安全性要求不高的场合可以采用RAID5方式来折中；其他应用中采用大容量的ATA或SATA硬盘结合raid10，既降低了raid10的为获得一定的存储空间必须采用双倍磁盘空间的拥有成本，又避免了raid5相对raid10的各种缺点。

在企业应用中，raid10结合SATA磁盘意味着一个更好的选择。

第5章 主机系统高可用技术

5.1 概述

随着计算机和网络的飞速发展，计算机在各个行业的应用越来越广泛和深入。在绝大多数行业、绝大多数企业都存在一些关键的应用，这些应用必须7*24*365小时不间断运行。这些应用的主机系统一旦出现问题，轻则降低业务响应速度，严重的会导致业务中断，造成严

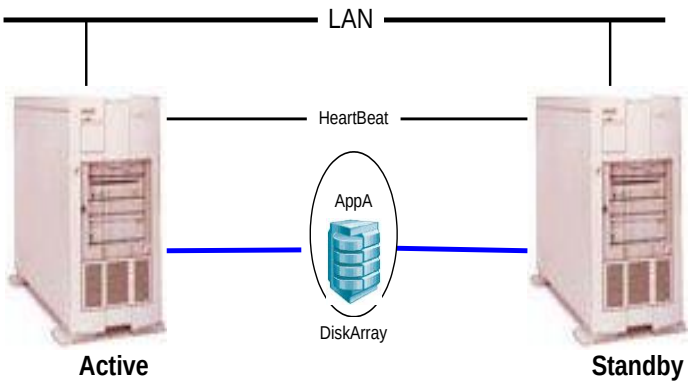
重的后果。如何能保证业务的持续进行，已经成为影响一个公司成败的关键因素。在这样的情况下，系统的高可用性就显得尤为重要。

近年来，服务器平台的可用性在不断地提高。内存ECC(错误代码校正)及Chipkill技术（纠正及探测内存中的数据错误）、硬盘RAID技术、网络负载均衡及容错技术以及多种基于硬件的冗余设计（如硬盘子系统、风扇子系统、电源子系统等）提高了整个系统的可用性，较好的保证了业务系统的持续运行。虽然硬件技术的发展大大提高了系统的可靠性，但是，由于系统内其它核心部件（如CPU、主板、物理内存等）的故障， 应用系统在一年365天内还是可能出现44～87小时的停机时间，这就要求从更高层次、更多方面综合考虑提高系统的高可用性。

在高可用技术中，根据不同的应用环境，从性能、经济等方面考虑.主要有以下几种方法和模式：

5.1.1 双机热备份方式

在双机热备份方式中，主服务器运行应用，备份服务器处于空闲状态，但实时监测主服务器的运行状态。一但主服务器出现异常或故障，备份服务器立刻接管主服务器的应用。也就是目前通常所说的 active/standby 方式，主要通过纯软件方式实现双机容错。



当前应用最广泛的双机热备份软件主要有 LifeKeeper,Rose HA, DataWare 和 MSCS。

LifeKeeper 双机软件：

LifeKeeper 双机软件是美国 SteelEye 公司的产品，支持 Windows NT(2000),Linux,UNIX 等平台操作系统，主要提供数据、应用程序的可用性保护。

LifeKeeper 可以支持二到十六个服务器结点，在出现故障的情况,LifeKeeper 会将保护资源自动转换到一个根据优先权而设定的主机系统上。

产品特点：

- 支持 Notes、Exchange、SQL Server、Sybase、Informix、Oracle 、SAP 等多种应用系统
- 支持 NT、Windows 2000、Linux 和 NCR Unix 操作系统
- 使用共享磁盘阵列柜方式时，最多可以支持 16 个节点

LifeKeeper 的产品版本比较复杂，以 Windows 环境为例，实现数据库的双机热备软件包括 Lifekeeper for Windows，MS SQL 2000 ARK 等。前者实现操作系统和存储级的切换，但如果要实现相关应用的切换，则还需要同时安装相应的 ARK。

下表为 LifeKeeper 的部分常用产品规格及参考价格：

产品编	产品规格描述	列表价	市场	实际

号		格	参考价格	出货价
2374-2	LifeKeeper for Windows 2000 v4.2	$3750 x 2	28000	7500 元
2390	Microsoft SQL2000 ARK for Windows 2000 v4.0	$1200	4000	1300 元

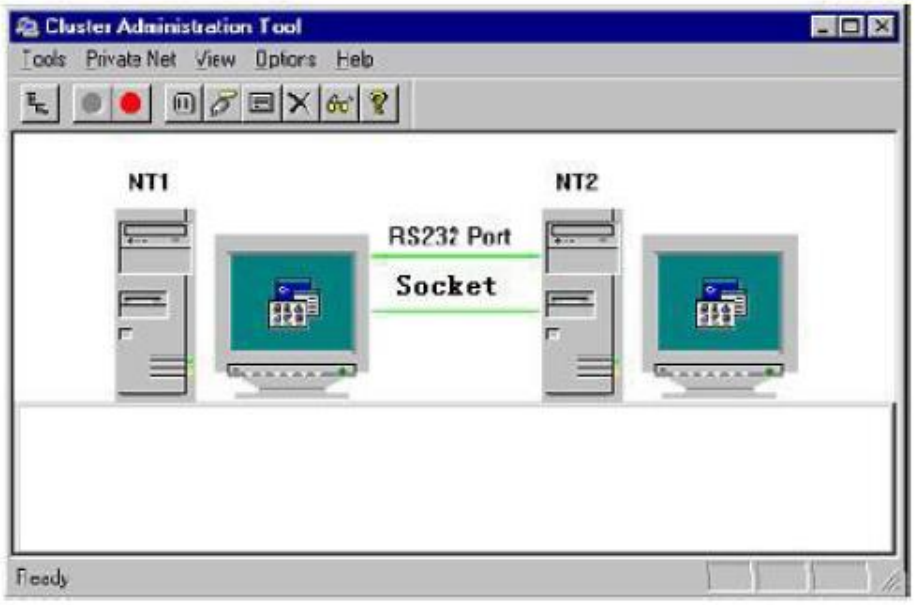
Rose HA 双机软件：

Rose HA 双机软件是美国 Rose 数据公司的产品，支持 Windows NT/2000/2003、SCO Openserver/Unixware、Linux、Solaris 操作系统平台，支持 MS-SQL、Oracle 、Informix、Sybase、Lotus/Nose、DB2 等数据库。接管动作包括：文件系统(File System)、数据库(Database)、 网络地址(IP Address) 、应用程序(AP) 以及系统环境(OS)。

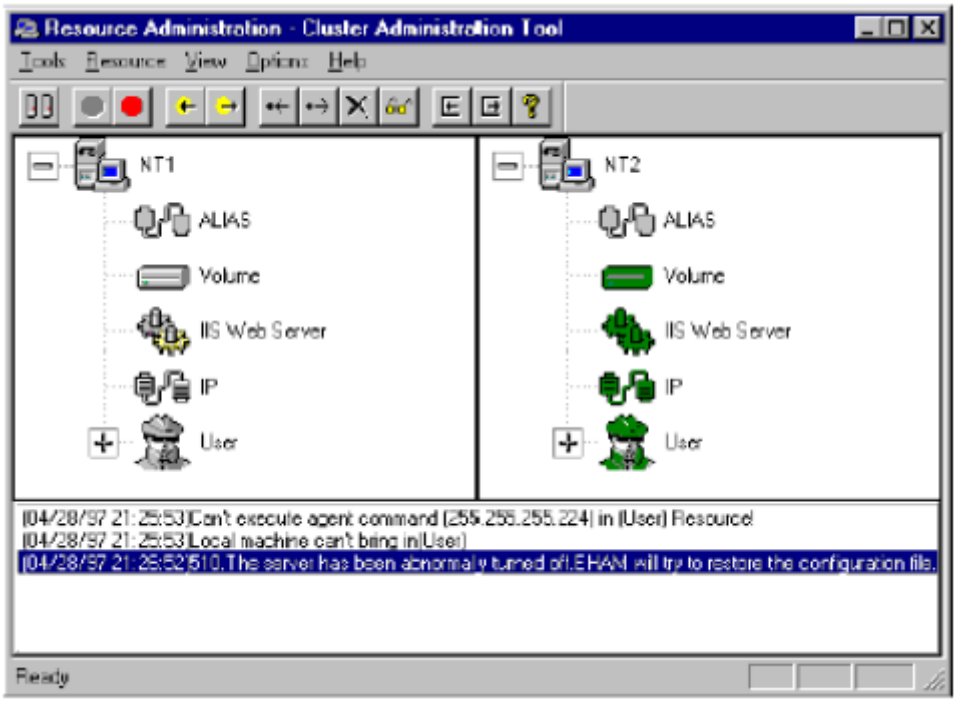
产品特点：

- 提供两节点集群高可用性方案
- 纯软件解决方案，针对系统不可避免的预期和非预期宕机问题
- 系统宕机或服务中断时可提供错误识别、故障分离和服务的在线恢复等功能
- 支持对多种常用系统服务和应用软件的保护
- 通过用户自定义资源管理，支持其它应用软件的保护

图形管理界面



资源项目的图形管理界面



工作模式：如上图所示

1) 双主机通过一条 TCP/IP 网络线以及一条 RS-232 电缆线相联
2) 双主机各自通过一条 SCSI 电缆线与 RAID 相联
3) 主机 NT1 为 active，主机 NT2 为 standby
4) 主机 NT1 处理作业和数据，主机 NT2 作为热备份机
5) 主机 NT1 故障后，主机 NT2 自动接管主机 NT1 的作业和数据
6) 主机 NT2 同时接管 NT1 的主机名(Host)及网络地址(IP)
7) 主机 NT1 的作业将在主机 NT2 上自动运行
8) 主机 NT1 的客户(client)可继续运行，无需重新登录
9) 主机 NT1 修复后，自动接管原来的作业和数据，主机 NT2 继续作备份机

下表为Rose HA的部分常用产品规格及参考价格：

产品编号	产品规格描述	列表价格	市场参考价格	出货价格参考	备注
-	Rose HA v6.0	36000	18000	9100元	一套软件支持两台服务器。支持Win 2000 /NT/2003 Server及SQL Server

DataWare 双机软件：

DataWare 双机软件是台湾豪威科技公司的产品，支持基于 Windows NT 平台和 Unix 平台的操作系统。

产品特点：

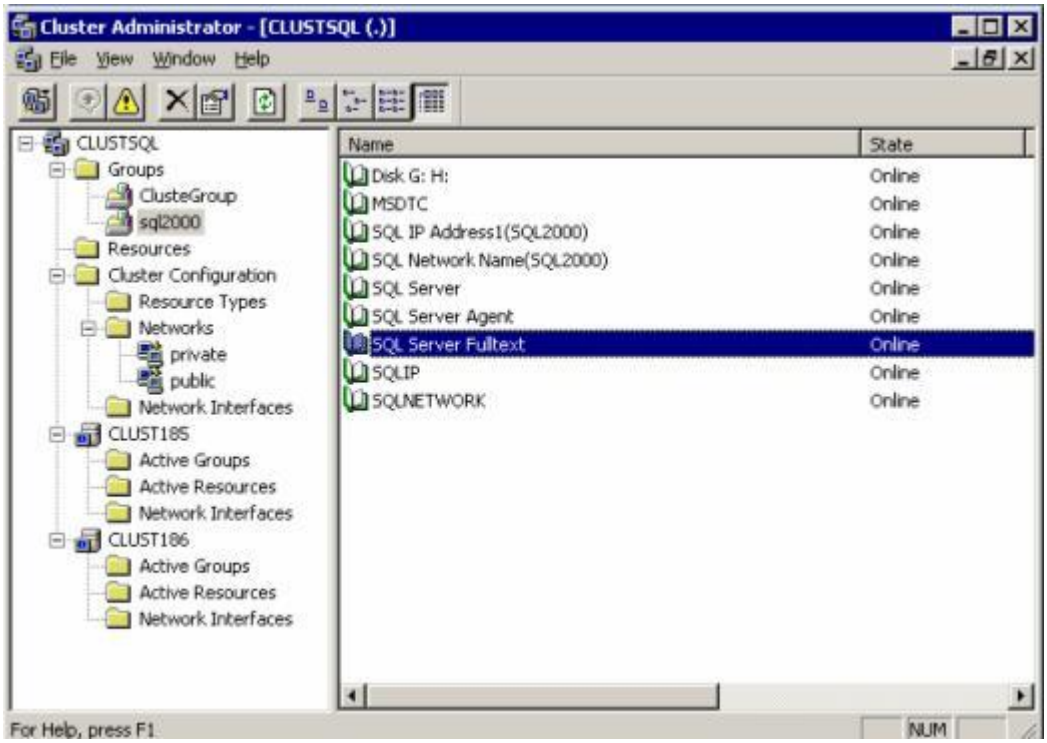
- 应用重启：任何造成工作中断的故障首先试图在本地解决(重新在本服务器启动)。如果不可恢复，所有资源将在最短时间内在对方服务器上启动。
- 自动客户端连接：当应用切换后，你不需要通知和重新配置客户端。
- 自动事件触发：系统管理员可定义脚本文件，DataWare 检测到相应事件后将自动启动脚本文件。
- 自动或手动切换：除了自动切换功能外，DataWare 还提供手动切换功能。这种特性可以使用户在需要做定期维护或系统升级时不会造成客户端应用长时间的停顿。

下表为DataWare的部分常用产品规格及参考价格：

产品编号	产品规格描述	列表价格	市场参考价格	出货价格参考	备注
-	DataWare	38000	14000	8000元	一套软件支持两台服务器。

MSCS 双机软件：

MSCS 双机软件是 Microsoft 的产品，所采用的容错软件--Microsoft Cluster System（MSCS）软件集成在 WIN NT Enterprise server /WIN 2000 Advance Server 中。对不少的应用程序资源提供了保护，在 WIN NT/WIN 2000 的服务列表中的所有服务全部可以提供保护。支持 MS SQL Server、Oracle、Sybase、Informix 等数据库和支持 Notes、Web 服务器、FTP 服务器、MS Exchange、SAP 等应用软件。

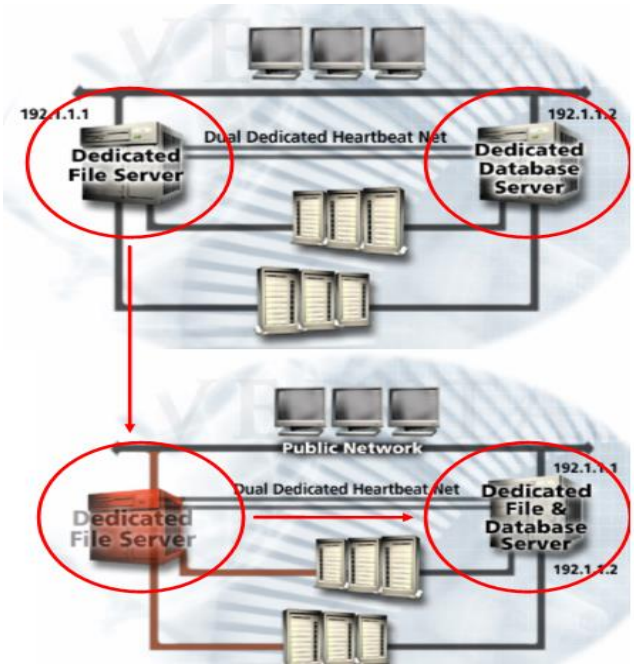


产品特点：

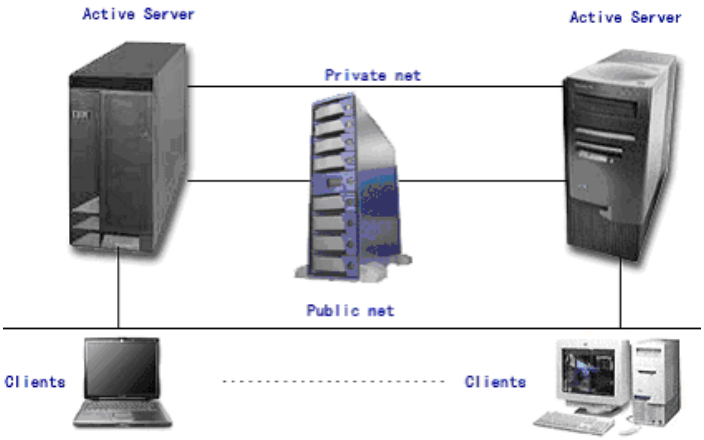
- 应用级的高性能切换：可以实现系统级的服务器切换，而且提供强大的应用级服务器切换，表现在对任意应用可以进行检测并可以分为不同的资源组切换到不同的服务器
- 易管理性、易使用性：MSCS 系统安装简单，易于维护。占用系统资源极少，不增加网络负荷，且不打扰任何具体应用系统的任何操作。图形界面操作，简单方便。
- 多种配置实现：可以实现双机直接连接也可以实现基于 SAN 的全冗余结构。

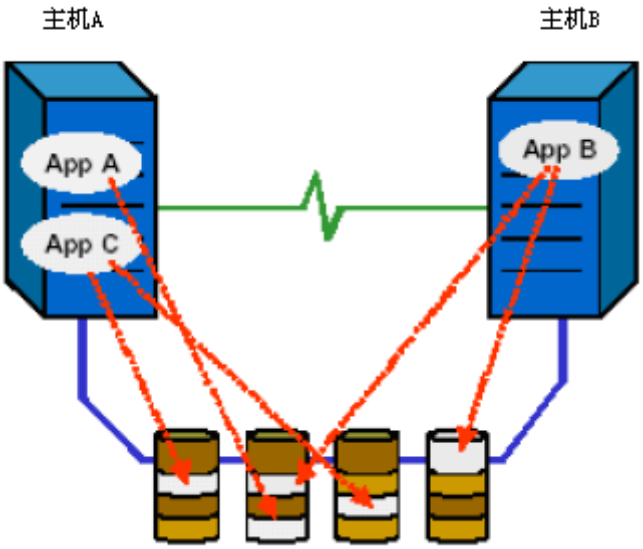
5.1.2 双机互备份方式

在这种方式中，没有主服务器和备份服务器之分，两台主机互为备份。主机各自运行不同应用，同时还相互监测对方状况。当任一台主机宕机时，另一台主机立即接管它的应用，以保证业务的不间断运行。也就是目前通常所说的 Active/Active 方式，主要通过纯软件方式实现双机容错。通常情况下，支持双机热备的软件都可以支持双机互备份方式，当前应用最广泛的双机互备软件主要有 LifeKeeper，Rose HA，DataWare 和 MSCS。


192.1.1.1
Dual Dedicated Heartbeat Net
192.1.1.2
Dedicated File Server
Dedicated Database Server
Public Network
Dual Dedicated Heartbeat Net
Dedicated File Server
Dedicated File & Database Server
192.1.1.2

以 Rose 为例：


Active Server
Active Server
Private net
Public net
Clients
Clients



1) 双主机通过一条 TCP/IP 网络线以及一条 RS-232 电缆线相联
2) 双主机各自通过一条 SCSI 电缆线与 RAID 磁盘阵列相联
3) 双主机各自运行不同的作业，彼此独立，并相互备援
4) 主机 A 故障后，主机 B 自动接管主机 A 运行
5) 主机 A 的作业将在主机 B 上自动运行
6) 主机 A 的客户(client)要在主机 B 上重新登录
7) 主机 A 修复后，主机 B 将把 A 的作业自动交还主机 A
8) 主机 B 故障时，主机 A 接管主机 B 的作业和数据

主机 B 修复时,主机 A 再将原来接管的作业和数据交还主机 B。

5.1.3 群集并发存取方式

在这种方式下，多台主机一起工作，各自运行一个或几个服务。当某个主机发生故障时，运行在其上的服务就被其它主机接管。群集并发存取方式在获得高可用性的同时，也显著提高了系统整体的性能。主要的群集软件有集成了 Windows 群集（Windows Clustering）软件的 Microsoft® Windows® Server 2003 Enterprise Edition，Veritas 的 cluster server 和一些基于 Linux 开发的集群管理软件,一般都支持八个以上节点的群集。

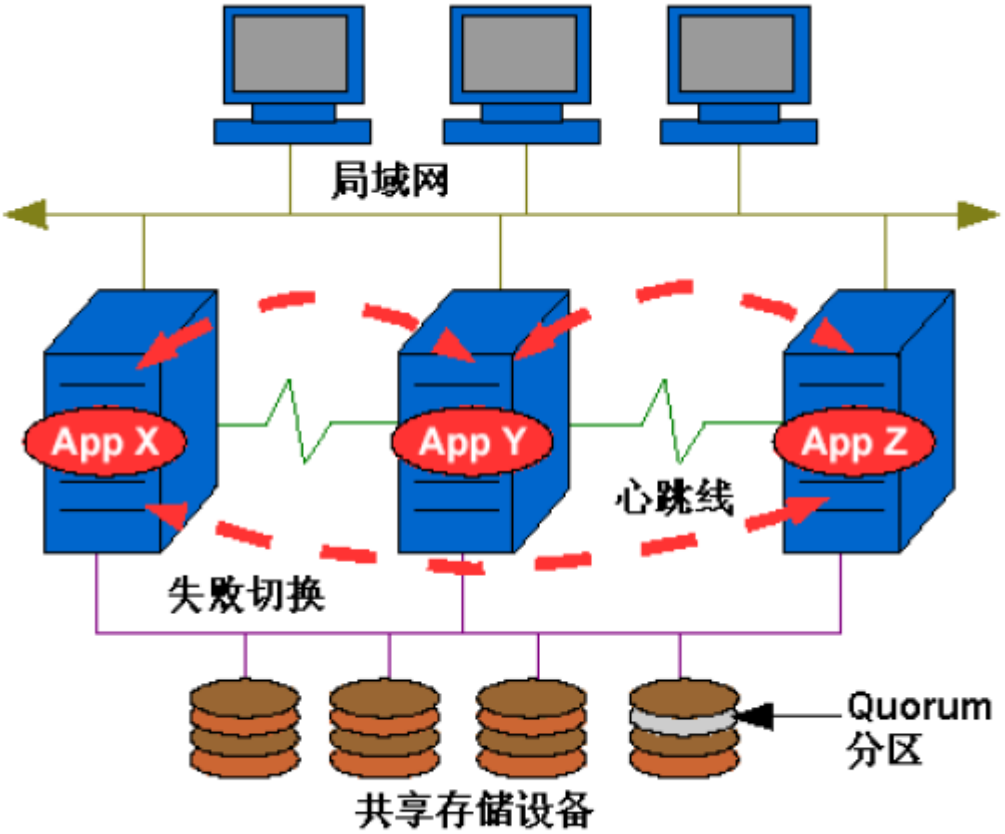
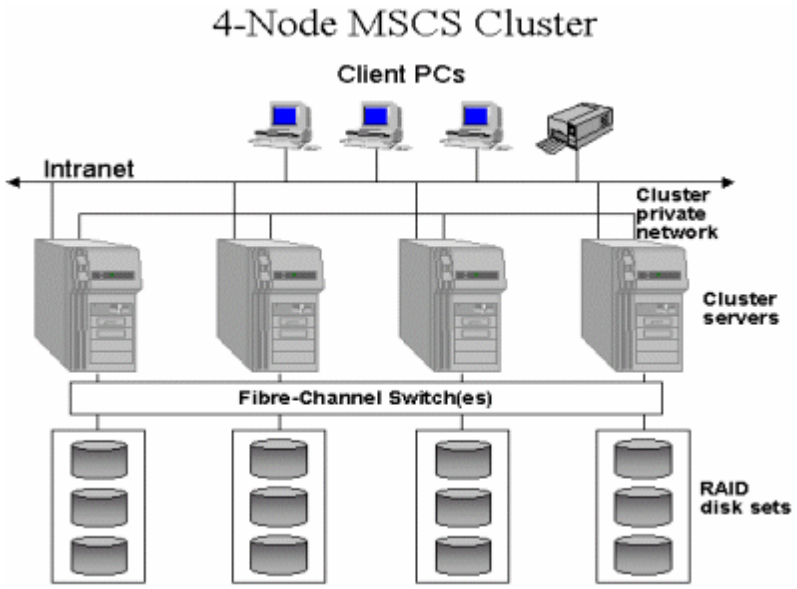


图 1 – 典型的高可用群集配置

以Windows群集（Windows Clustering）软件为例，下图简要描绘了一个四节点群集的配置：



如果群集中的某一台服务器由于故障或维护需要而无法使用，资源和应用程序将转移到可用的群集节点上。能够为多数关键任务应用程序提供足够的可用性。群集服务可以对应用程序和资源进行监控。当群集应用程序的总体负荷超出了群集的能力范围时，可以添加附加的节点，来满足需求的增长。

5.2 工作模式

5.2.1 双机热备份方式

所谓双机热备份就是一台主机为主服务器（Active Server），另一台主机为备份服务器（Standy Server），在系统正常情况下，主服务器为应用系统提供支持，备份服务器监视主服务器的运行情况。当主服务器出现异常，不能支持应用系统运行时，备份机主动接管工作机的工作，继续支持应用系统的运行，从而保证信息系统能够不间断的运行。此时，原来的备份服务器就成了主服务器。当原来的主服务器经过修复正常后，系统管理员通过管理命令或经由以人工或自动的方式将备份服务器的工作切换回主服务器；也可以激活监视程序，监视备份服务器的运行情况。在正常情况下，主服务器也会监视备份机的状态，当备份服务器因某种原因出现异常时，工作服务器会发出告警，提醒系统管理员解决故障，以确保主/备服务器切换的可靠性。

5.2.2 双机互备方式

所谓双机互备就是两台服务器均为工作机，在正常情况下，两台工作机均为应用系统提供支持，并互相监视对方的运行情况。当一台主机出现异常，不能支持应用系统正常运营时，另一主机则主动接管异常机的应用，从而保证应用系统能够不间断的运行。但是，当一台主机出现异常并被接管后，正常运行的主机的负载会随之加大，严重的情况下有可能影响到应用系统的响应速度。所以此时必须尽快修复异常机，以缩短正常机单机运行的时间。

5.2.3 群集并发存取方式

所谓群集（Cluster）技术就是一个域内包含多台拥有共享存储空间的服务器，各服务器通过内部局域网相互通信，群集内的任一服务器上运行的业务都可被所有的客户所使用。当一台服务器发生故障时，它所运行的应用将由其他服务器自动接管，这就实现了负载均衡和

互为备份。

5.3 适用场合

三种工作方式，各自适合的应用场合。

- 双机热备方式：适用于硬件资源充足，对应用系统有严格高可靠性要求的企业、政府、军队、重要商业网站 ISP/ICP 或数据库应用等用户。这些用户不仅保证主机系统能够 24 小时提供不间断的服务，还要求发生故障切换时，应用系统的性能和响应速度不受影响，以确保网络系统、网络服务、共享磁盘空间、共享文件系统、进程以及数据库的高速持续运转。
- 双机互备方式：适用于在确保应用不间断运行的前提下，从投资的角度考虑，能充分的利用现有的硬件资源的用户。这些用户的应用要求保证业务不间断运行，但在发生故障切换时，允许一定时间内的应用性能的降低。
- 群集并发存取方式：适用于对计算数据处理要求高的应用，其特点是实时性强、阶段性数据流量大、对应用系统有严格高可靠性要求。这种方式需要更多的硬件投资，为企业带来更大的可靠性和更多的任务能力。

和前面提到的两种的高可用的计算机技术相比，群集技术并不要求所有服务器的性能相当，不同档次的服务器都可以作为群集的节点。在需要运行高负载的应用任务时，可以通过临时接入新的节点的方法，增加系统的运算和响应能力。群集技术系统可以在低成本的条件下完成大运算量的计算，具有较高的运算速度和响应能力，能够满足当今日益增长的信息服务的需求。群集技术适用于以下场合：

- 大规模计算如基因数据的分析、气象预报、石油勘探需要极高的计算性能。
- 应用规模的发展使单个服务器难以承担负载。
- 不断增长的需求需要硬件有灵活的可扩展性。
- 关键性的业务需要可靠的容错机制。

5.4 对存储系统的要求

- 双机热备方式：系统运行时，只有主服务器与存储系统进行数据交换。当发生主机故障切换时，要求存储系统能与备份服务器快速建立数据通道，以支持业务的快速切换。
- 双机互备方式：系统运行时，两台主机需要同时对磁盘阵列进行读写操作，这要求存储系统具备良好的的并发读取操作和一定的负载均衡功能。
- 群集并发存取方式：

1、 并发处理能力

高性能群集主要依赖高性能存储以满足其强大的运算能力和数据的读写运算，但多个群集节点的数据访问是并发的、无规律的，因此就要求存储设备具有很强的处理并发数据访问能力，以使群集应用发挥最高的性能。

2、 数据共享能力

高性能群集主要利用分布在多个节点的处理器共同计算存储系统里的数据。这就对存储系统的初始容量、后期容量扩充能力提出了很高的要求。同时，多个节点的处理器

能够方便地共享相关的数据，这就要求存储系统具备安全而高效的共享能力。

3、 大规模与可扩展性

随着高性能群集系统内计算节点的数量与规模、每个网络的数据容量也在扩大。因此，中央存储系统是否具备方便的升级途径和巨大的可供升级容量，就成为重要的因素。如何实现在线升级、平滑过渡、现有用户及素材的透明化处理，是存储产品必需的功能。

4、 可管理性

一是管理操作分安全级别；二是提供清晰明确的管理界面，方便操作。避免人为误操作，要求存储系统的管理界面简单明了，管理操作流程设计合理。

5、 高可用性

高性能群集的时效性很强，因此要求网络系统具有极高的可靠性。但是绝对的安全性是没有的，必要的网络故障恢复时间就显得十分重要。首先要求有较高的容错级别，例如控制器要求高可用容错，存储子系统要求容错冗余等；其次故障恢复时间要短，尽可能做到不宕机的在线恢复。

第6章 数据一致性

6.1 数据一致性概述

数据一致性是指关联数据之间的逻辑关系是否正确和完整。问题可以理解为应用程序自己认为的数据状态与最终写入到磁盘中的数据状态是否一致。比如一个事务操作，实际发出了五个写操作，当系统把前面三个写操作的数据成功写入磁盘以后，系统突然故障，导致后面两个写操作没有写入磁盘中。此时应用程序和磁盘对数据状态的理解就不一致。当系统恢复以后，数据库程序重新从磁盘中读出数据时，就会发现数据再逻辑上存在问题，数据不可用。

6.2 Cache 引起的数据一致性问题

引起数据一致性问题的一个主要原因是位于数据I/O路径上的各种Cache或Buffer（包括数据库Cache、文件系统Cache、存储控制器Cache、磁盘Cache等）。由于不同系统模块处理数据IO的速度是存在差异的，所以就需要添加Cache来缓存IO操作，适配不同模块的处理速度。这些Cache在提高系统处理性能的同时，也可能会“滞留”IO操作，带来一些负面影响。如果在系统发生故障时，仍有部分IO“滞留”在IO操作中，真正写到磁盘中的数据就会少于应用程序实际写出的数据，造成数据的不一致。当系统恢复时，直接从硬盘中读出的数据可能存在逻辑错误，导致应用无法启动。尽管一些数据库系统（如Oracle、DB2）可以根据redo日志重新生成数据，修复逻辑错误，但这个过程是非常耗时的，而且也不一定每次都能成功。对于一些功能相对较弱的数据库（如SQL Server），这个问题就更加严重了。

解决此类文件的方法有两个，关闭Cache或创建快照（Snapshot）。尽管关闭Cache会导致系统处理性能的下降，但在有些应用中，这却是唯一的选择。比如一些高等级的容灾方案中（RPO为0），都是利用同步镜像技术在生产中心和灾备中心之间实时同步复制数据。由于数据是实时复制的，所以就必须要关闭Cache。

快照的目的是为数据卷创建一个在特定时间点的状态视图，通过这个视图只可以看到数据卷在创建时刻的数据，在此时间点之后源数据卷的更新（有新的数据写入），不会反映在快照视图中。利用这个快照视图，就可以做数据的备份或复制。那么快照视图的数据一致性

是如何保证的呢？这涉及到多个实体（存储控制器和安装在主机上的快照代理）和一系列的动作。典型的操作流程是：存储控制器要为某个数据卷创建快照时，通知快照代理；快照代理收到通知后，通知应用程序暂停IO操作（进入backup模式），并flush数据库和文件系统中的Cache，之后给存储控制器返回消息，指示已可以创建快照；存储控制器收到快照代理返回的指示消息后，立即创建快照视图，并通知快照代理快照创建完毕；快照代理通知应用程序正常运行。由于应用程序暂停了IO操作，并且flush了主机中的Cache，所以也就保证了数据的一致性。

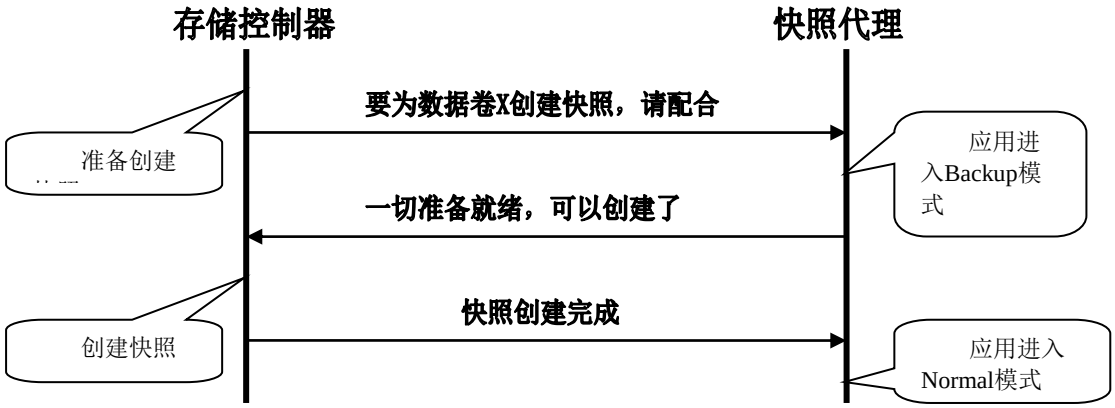


图6-1 快照的创建过程

创建快照是对应用性能是有一定的影响的（以Oracle数据库为例，进入Backup模式大约需要2分钟，退出Backup模式需要1分钟，再加上通信所需时间，一次快照需要约4分钟的时间），所以快照的创建不能太频繁。

6.3 时间不同步引起的数据一致性问题

引起数据不一致性的另外一个主要原因是对相关联的多个数据卷进行操作（如备份、复制）时，在时间上不同步。比如一个Oracle数据库的数据库文件、Redo日志文件、归档日志文件分别存储在不同的卷上，如果在备份或复制的时候未考虑几个卷之间的关联，分别对一个个卷进行操作，那么备份或复制生成的卷就一定存在数据不一致问题。

此类问题的解决方法就是建立“卷组（Volume Group）”，把多个关联数据卷组成一个组，在创建快照时同时为组内多个卷建立快照，保证这些快照在时间上的同步。之后再利用卷的快照视图进行复制或备份等操作，由此产生的数据副本就严格保证了数据的一致性。

6.4 文件共享中的数据一致性问题

通常所采用的双机或集群方式实现同构和异构服务器、工作站与存储设备间的数据共享，主要应用在非线性编辑等需要多台主机同时对一个磁盘分区进行读写。

在NAS环境中，可以通过网络共享协议N FS或CIFS来做到数据的共享。但是如果不在NAS环境中，多台主机同时对一个磁盘分区进行读写会带来写入数据一致性的问题，造成文件系统被破坏或者当前主机写入后其它主机不能读取当前写入数据的问题。

可以通过使用数据共享软件装在多台主机上来实现磁盘分区的共享。由数据共享软件来调配多台主机数据的写入，保证数据的一致性。

第7章 数据复制与容灾

7.1 灾难恢复/业务连续性

随着企业信息化进程的不断深化，信息系统成为了支撑企业业务运行的重要平台，企业的全部业务流程都依赖于信息系统提供的服务来运作。这种统一的业务运作平台在简化业务流程，提高工作效率的同时，也带来了安全性方面的全新要求。那就是信息系统必须具备抵抗灾难的能力，具备在灾后快速恢复的能力，只有这样，才能满足企业业务连续性的需求。

在国内，尽管企业对信息系统的重要性和容灾需求早有认识，但鉴于适用技术、方案成本等多方面原因，容灾系统的建设一直属于企业的自主行为。在9.11事件和印度洋海啸之后，国家充分认识到了重要信息系统容灾的必要性，要求一些重要行业的信息系统必须实现容灾。为了加强对信息系统安全的管理，规范对信息系统灾难性故障的响应和处置，国务院信息化办公室在2005年发布了《重要信息系统灾难恢复指南》，用于指导信息系统的使用和管理单位的灾难恢复规划工作，以及对信息系统灾难恢复规划项目的审批和监督管理。

《指南》给灾难下了一个清晰的定义，即“由于人为或自然的原因，造成信息系统运行严重故障或瘫痪，使信息系统支持的业务功能停顿或服务水平不可接受、达到特定的时间的突发性事件”。这个定义不仅给出了灾难的范围，也给出了灾难的判断标准。灾难不只包括自然灾难（地震、海啸等），也包括人为的灾难（如恐怖袭击、误操作、病毒等）。另外，判断信息系统是否因灾难而故障的标准除“支持的业务停顿”之外，“服务水平不可接受”也是一个方面。事实上，“服务水平不可接受”是比较难以定量的分析判定的，这增加了实现自动化故障切换（如“零秒”切换）的难度。

参照国际相关标准，并结合国内实际情况，《指南》还将灾难恢复应具有的技术和管理支持分为6个等级，每个级别都包括数据备份系统、备用数据处理系统、备用网络系统、备用基础设施、技术支持、运行维护支持和灾难恢复预案这7个要素。在7个要素中，前三个属于IT技术的范畴，而后四个属于管理和服务的范畴。其中，数据备份系统面向的对象是数据，

目的是实现数据的冗余备份，以便一份数据被破坏以后，还有另外一份数据可用，常用的技术有数据备份（Backup）和数据复制（Replication）等。备用数据处理系统面向的对象是应用服务器，目的是在主用数据处理系统发生故障以后，可以利用数据备份系统产生的冗余数据来恢复应用，常用的技术有服务器双机热备、服务器集群等。备用网络系统面向的是网络连接，目的是保证备用数据处理系统与其客户端、不同备用数据处理系统之间的网络，以便整个实现业务系统的恢复。

要素 / 级别	数据备份系统	备用数据处理系统	备用网络系统	备用基础设施	技术支持	运行维护支持	灾难恢复预案
级别1 级别2 级别3 级别4 级别5							
级别6							

七个要素的不同满足程度决定了容灾方案的等级，等级的划分最终反映在技术指标上，不同等级的容灾方案对应有不同的技术指标值。常用的容灾方案评价指标主要有RTO（Recovery Time Object，恢复时间目标）、RPO（Recovery Point Time，恢复点目标）和容灾半径。RTO是指“将信息系统从灾难造成的故障或瘫痪状态恢复到可正常运行状态，并将其支持的业务功能从灾难造成的不正常状态恢复到可接受状态”所需时间，其中包括备份数据恢复到可用状态所需时间、数据处理系统切换时间、以及备用网络切换时间等，该指标用以衡量容灾方案的业务恢复能力。RPO是指业务系统所允许的灾难过程中的最大数据丢失量（以时间来度量），这是一个与数据备份系统所选用的技术有密切关系的指标，用以衡量灾难恢复方案的数据冗余备份能力。容灾半径是指生产中心和灾备中心之间的直线距离，用以衡量容灾方案所能防御的灾难影响范围。显然，具有零RTO、零RPO和大容灾半径的灾难恢复方案是用户最期望的，但受系统性能要求、适用技术及成本等方面的约束，这种方案实际上是不大可行的。所以，用户在选择容灾方案时应该综合考虑灾难的发生概率、灾难对数据

的破坏力、数据所支撑业务的重要性、适用的技术措施及自身所能承受的成本等多种因素，理性地作出选择。

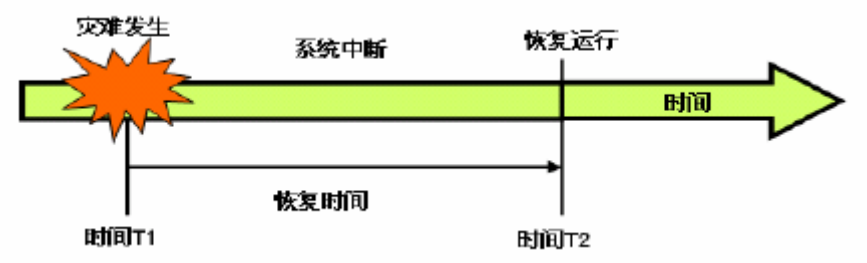


图1 RTO 示意图

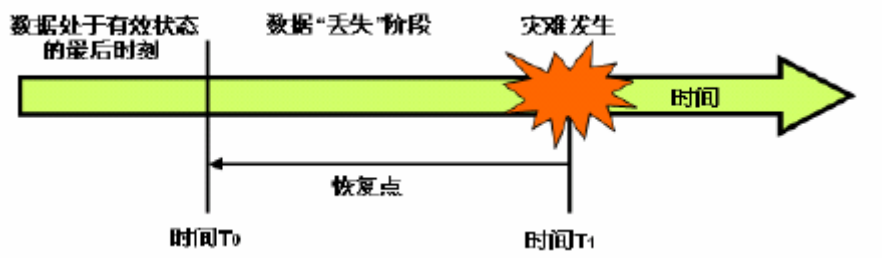


图2 RPO 示意图

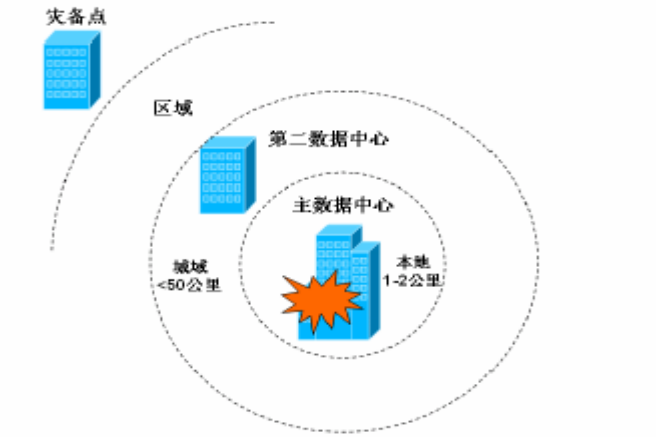


图3 灾难半径示意图

除技术指标以外，容灾方案的ROI（Return of Investment，投入产出比）也是用户需要重点关注的，它用以衡量用户投入到容灾系统的资金与从中所获得的收益的比率。表明上看，容灾系统不像其它业务系统那样会给用户带来收益，但事实上，容灾系统确实是有收益的，

而且收益是完全可以度量的。容灾系统的收益主要来源于发生灾难时为用户所挽回的损失，这种损失不只包括金钱方面的，信誉、客户忠诚度、法律风险等方面的损失也包含在内。业界统计数据表明，随着业务停运时间的延长，用户的损失会急剧增加。当然不同行业的损失程度有所不同，其中以金融、电信为最。如果容灾系统能够把由于灾难而导致的业务停运时间显著缩短，也就间接为客户创造了收益。基于容灾方案的技术指标、业界的统计数据和用户自身业务状况，用户是完全可以对容灾方案的收益作出一个适当的量化评估的。在ROI指标方面，基于新型IP SAN系统的容灾方案显得更有优势，因为这类方案不仅能大幅降低容灾系统的初始部署成本，而且管理成本也相对要低很多。

Cost of Downtime is High	
Industry	*Average Loss/Hr.*
Brokerage Operations	$6,450,000
Credit card Authorizations	$2,600,000
E-Commerce	$240,000
Package Shipping Services	$150,000
Home Shopping Channels	$113,000
Catalog Sales Center	$90,000
Airline Reservation Center	$88,000
Cellular Service Activation	$41,000
ATM Service Fees	$14,500

7.2 数据备份系统

在构建容灾系统所涉及的7个要素中，数据备份系统是基础，只有保证了数据的安全可用，业务的恢复才有可能。数据备份系统采用的技术主要有数据备份（Backup）和数据复制（Replication）两种。

7.2.1 数据备份

数据备份（Backup）一般是指利用备份软件（如Veritas的NetBackup、CA的BrightStor等）把数据从磁盘备份到磁带进行离线保存（最新的备份技术也支持磁盘到磁盘的备份，也就是把磁盘作为备份数据的存放介质，以加快数据的备份和恢复速度）。备份数据的格式是

磁带格式，不能被数据处理系统直接访问。在源数据被破坏或丢失时，备份数据必须由备份软件恢复成可用数据，才可让数据处理系统访问。

数据备份在一定程度上是可以保证数据安全的，但应用于容灾系统时却面临众多问题：

1）备份窗口

备份窗口是指应用所允许的完成数据备份作业时间。由于数据备份作业会导致应用主机的性能下降，甚至服务水平不可接受，备份作业必须在应用停机或业务量较小的时候进行。但随着备份数据量的不断增加和业务7×24小时连续运行需求的提出，备份窗口的问题越来越突出。问题的解决之道主要在于加快备份速度（如采用高速带库、磁盘备份）和实现在线备份。

2）恢复时间

在容灾系统中，备份数据的恢复时间直接关系到容灾方案的RTO指标。当备份数据量较大或者备份策略比较复杂时，备份数据往往需要较长的恢复时间。

3）备份间隔

鉴于备份作业对主机系统的影响，两次备份作业之间的间隔不能太密集。以常用的备份策略（1个全备＋6个增量备份）为例，备份间隔为1天。也就是说如果在两次备份之间发生灾难，RPO（数据的丢失量）接近于1天，这对于一些重要的信息系统是完全不可接受的。

4）数据的可恢复性

数据备份的目的就是为了数据恢复。但往往由于介质失效、认为错误、备份过程出错等原因，造成备份数据的不可恢复。

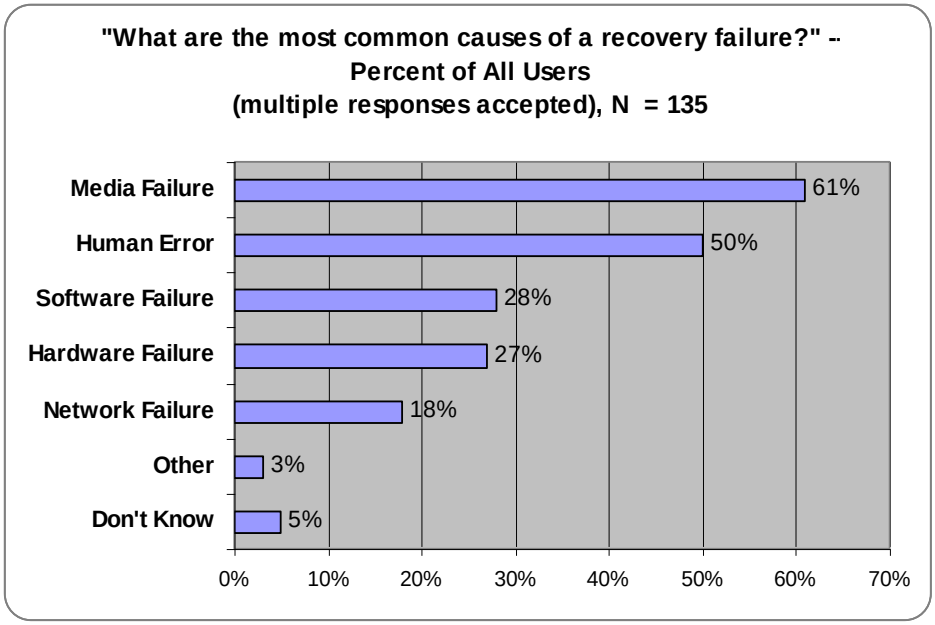


图4 备份数据恢复失败原因调查（数据来源：ESG）

5）介质的保管和运送

在完成数据备份以后，为了保证备份数据的安全性，一般采用的方式是把备份介质运输到远程的数据中心进行保管。但是在运输过程中，可能会造成备份数据的丢失。最近爆出的美国银行丢失120万名客户资料的事件就是佐证。

6）备份的成本

从提高备份速度和恢复速度，提高数据可恢复性方面来看，D2D是个不错的选择，但是现有备份软件的D2D选件都非常昂贵，方案实施成本比较高。

综合以上分析可以知道，高等级的容灾方案不适合于采用数据备份（Backup）技术来保证数据安全，数据备份只适合于一些低等级的容灾方案，对RTO和RPO要求相对比较低。但这并不意味这高等级容灾系统中不需要数据备份，作为一种廉价、成熟的技术，数据备份可以为容灾系统提供更多一层的保护。

7.2.2 数据复制

数据复制（Replication）是指利用复制软件（如EMC的SRDF、H3C同步异步镜像等）把数据从一个磁盘复制到另一个磁盘，生成一个数据副本。这个数据副本是数据处理系统直接可以访问的，不需要进行任何的数据恢复操作，这一点是复制与D2D备份的最大区别。

数据复制有多种分类方法，依据复制启动点的不同，数据复制可分为同步复制、异步复制、基于数据增量的复制等几种。对于同步复制，数据复制是在向主机返回写请求确认信号之前实时进行的；对于异步复制，数据复制是在向主机返回写请求确认信号之后实时进行的；而基于数据增量的复制是一种非实时的复制方式，它依据一定的策略（如设定数据变化量门限值、日历安排等）来启动数据复制。业界经常把不间断的，实时的数据复制称为镜像，所以同步/异步复制又被称为同步/异步镜像。

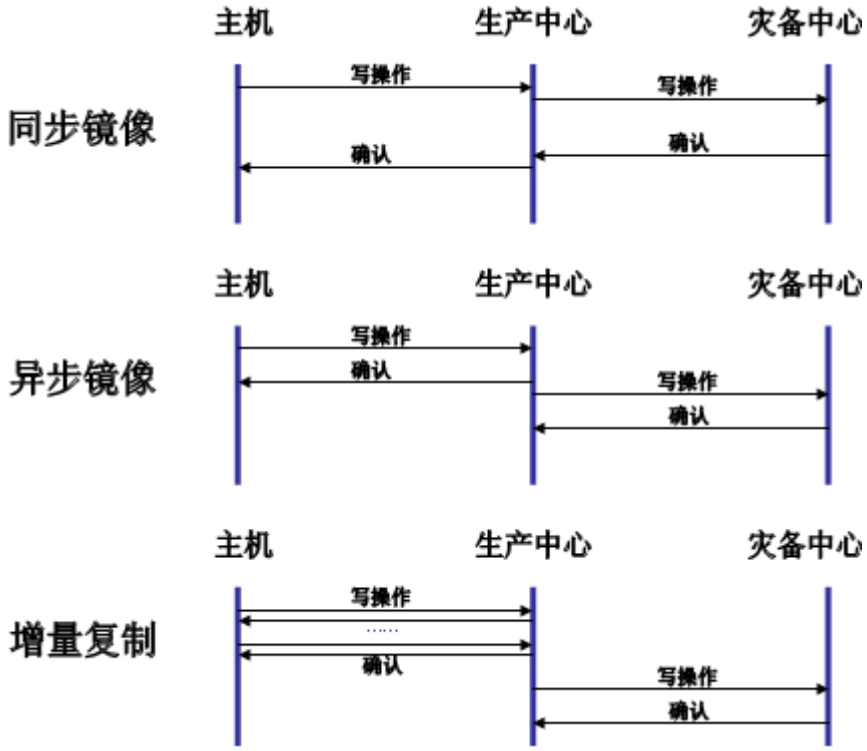


图5 同步镜像、异步镜像和增量复制

依据复制执行实体的不同，数据复制可分为基于主机的复制和基于存储设备的复制。基于主机的复制一般是由安装在主机中的软件插件来实施数据的复制，这会对主机系统的性能有所影响，典型的产品如Veritas的VVR，HP的OpenView SM等。基于存储设备的复制可以是由存储设备的控制器执行（如EMC的SRDF、华为3Com的同/异步镜像等），也可以是由虚拟化的存储管理平台来执行（如飞康IPStor的同/异步镜像和基于增量的复制）。基于存储设备的复制独立于主机平台，不会对主机系统的性能造成影响。

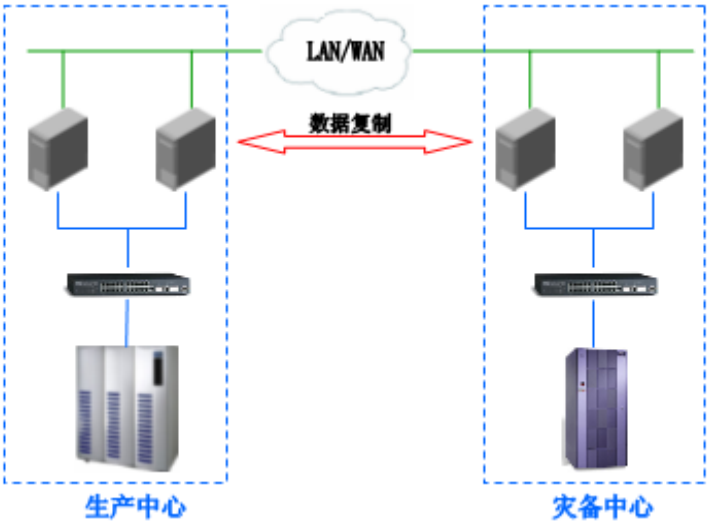


图6 基于主机的数据复制

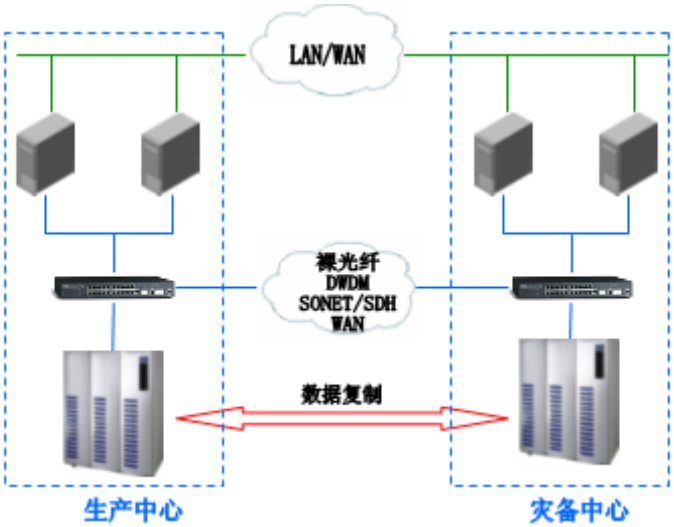


图7 基于存储设备的数据复制

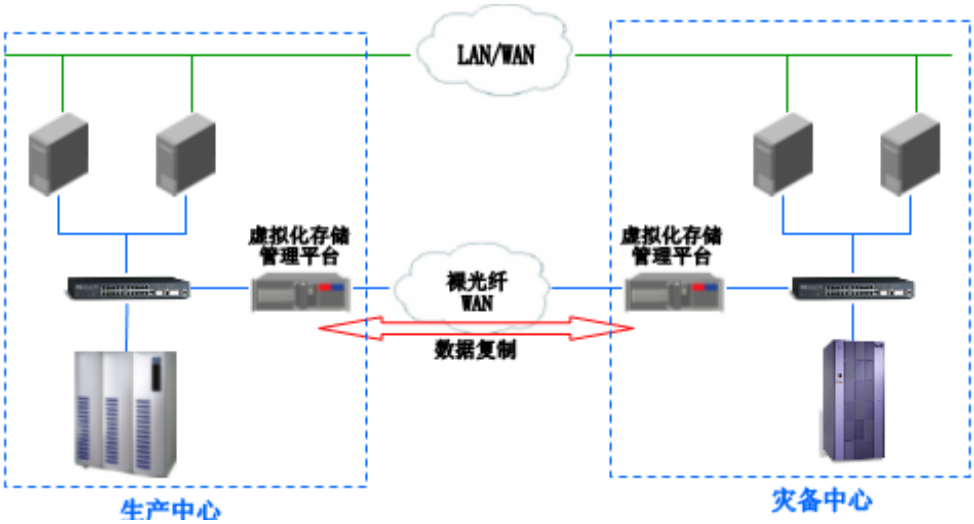


图8 基于虚拟化存储管理平台的数据复制

另外，依据数据复制站点之间的距离的不同，复制还可分为远程复制和本地复制。一般来说，复制距离小于1～2Km时为本地复制，大于该值时为远程复制。

7.3 数据一致性

在进行数据备份和数据复制时，保证数据的一致性是非常重要的。数据一致性问题可以理解为应用程序自己认为的数据状态与最终写入到磁盘中的数据状态是否一致。

引起数据一致性问题的主要原因是位于数据I/O路径上的各种Cache或Buffer（包括数据库Cache、文件系统Cache、存储控制器Cache、磁盘Cache等）。由于不同系统模块处理数据IO的速度是存在差异的，所以就需要添加Cache来缓存IO操作，适配不同模块的处理速度。这些Cache在提高系统处理性能的同时，也可能会“滞留”IO操作，带来一些负面影响。如果在系统发生故障时，仍有部分IO“滞留”在IO操作中，真正写到磁盘中的数据就会少于应用程序实际写出的数据，造成数据的不一致。当系统恢复时，直接从硬盘中读出的数据可能存在逻辑错误，导致应用无法启动。尽管一些数据库系统（如Oracle、DB2）可以根据redo日志重新生成数据，修复逻辑错误，但这个过程是非常耗时的，而且也不一定每次都能成功。对于一些功能相对较弱的数据库（如SQL Server），这个问题就更加严重了。

目前来说，数据一致性问题的解决之道有两个，关闭Cache或创建快照（Snapshot）。尽管关闭Cache会导致系统处理性能的下降，但在有些应用中，这却是唯一的选择。比如一些高等级的容灾方案中（RPO为0），都是利用同步镜像技术在生产中心和灾备中心之间实时同步复制数据。由于数据是实时复制的，所以就必须要关闭Cache。

快照的目的是为数据卷创建一个在特定时间点的状态视图，通过这个视图只可以看到数据卷在创建时刻的数据，在此时间点之后源数据卷的更新（有新的数据写入），不会反映在快照视图中。那么利用这个快照视图，就可以做数据的备份或复制。那么快照视图的数据一致性是如何保证的呢？这涉及到多个实体（存储控制器和安装在主机上的快照代理）和一系列的动作。典型的操作流程是：存储控制器要为某个数据卷创建快照时，通知快照代理；快

照代理收到通知后，通知应用程序暂停IO操作（进入backup模式），并flush数据库和文件系统中的Cache，之后给存储控制器返回消息，指示已可以创建快照；存储控制器收到快照代理返回的指示消息后，立即创建快照视图，并通知快照代理快照创建完毕；快照代理通知应用程序正常运行。由于应用程序暂停了IO操作，并且flush了主机中的Cache，所以也就保证了数据的一致性。

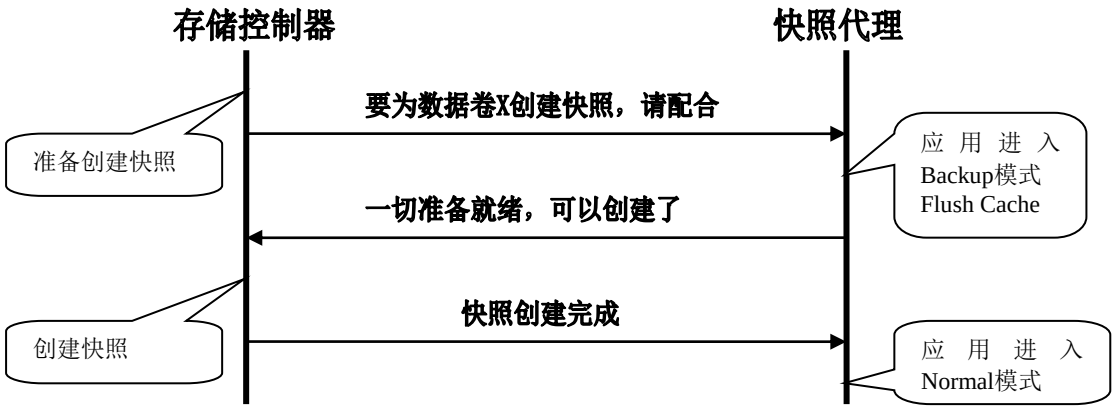


图9 快照的创建过程

创建快照是对应用性能是有一定的影响的（以Oracle数据库为例，进入Backup模式大约需要2分钟，退出Backup模式需要1分钟，再加上通信所需时间，一次快照需要约4分钟的时间），所以快照的创建不能太频繁。

7.4 总结

数据备份（Backup）：受备份策略、备份数据可恢复性等问题影响，不适合于在高等级的容灾方案（RPO小于24小时）作为构建备份数据系统主要的技术。但由于实施方便，成本低廉，适合于低等级的容灾方案中，也可作为高等级容灾方案的辅助技术。在应用停机的情况下进行数据备份就不存在数据一致性问题，当需要在线备份时，一般由备份软件来保证数据一致性。

同步镜像：应用于最高等级的容灾方案（RPO等于0）中，需要关闭主机Cache来保证数据一致性。对于连接生产中心和灾备中心的链路带宽和QoS要求很高，一般采用光纤直连、波分设备来保证，方案部署成本很高。

异步镜像：应用于较高级别的容灾方案（RPO接近于0）中，无法有效保证数据一致性（关闭主机中的Cache和快照都不适合）。但对于连接生产中心和灾备中心的链路带宽和QoS要求一般，理论上带宽只要达到“日新增数据量/（24×3600×8）”即可。

增量复制：应用于较高级别的容灾方案（RPO小于1小时）中，可以结合快照技术有效保证数据一致性。对于连接生产中心和灾备中心的链路带宽和QoS要求一般，理论上带宽只要达到“数据增量/复制间隔”即可。

第8章 备份技术

8.1 什么是备份

备份顾名思义，就是将数据以某种形式保存下来，备份的根本目的在于恢复，在这些数据丢失、毁坏和受到威胁的时候，使用数据的备份来恢复数据。虽然备份的定义可能很简单，不过具体实施存储系统的备份却可能是一份艰巨的任务，其中包含了许多可以预见的以及不易预见的需要考虑的因素。

8.2 备份与拷贝、归档的区别

备份不能仅仅通过拷贝完成，因为拷贝不能留下系统的注册表等信息；而且也不能留下历史记录保存下来，以做追踪；当数据量很大时，手工的拷贝工作又是何其麻烦。备份=拷贝+管理。管理包括备份的可计划性、磁带机的自动化操作、历史记录的保存以及日志记录等等。

正如生命周期理论将在线数据分级为在线和近线数据一样，离线数据亦可分为备份与存档数据，以降低投资和运维成本。

存档的目的是将需要长期备查或转移到异地保存/恢复的数据存放到可移动存储介质上。严格意义上讲，存档的目的不是为了保障数据安全，而只是为了实现数据仓储。如果说备份相当于桌头的字典，工作时会经常翻用，存档则好像日常工作中生成的一些具长期保存价值的文字资料，被转移到书架上或档案馆里备查。

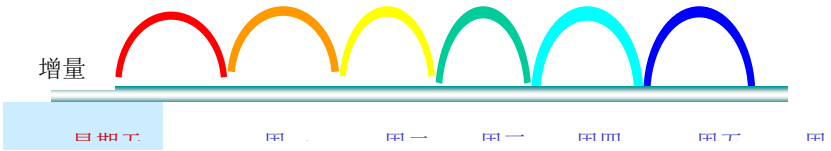
8.3 常规备份的实现方式

通常一套完整的备份系统包含备份软件、磁带机/磁带库、和备份服务器，具体的备份策略的制定、备份介质的管理以及一些扩展功能的实现，都是由备份软件来最终完成的。在备份服务器上安装备份软件的服务器端，在应用服务器端安装备份软件的客户端代理，如果是数据库应用还需要相应的数据库接口程序，客户端代理软件和服务器端软件协调工作，按照预先制定的备份策略自动或手动的将数据备份到磁带上。然而一个具有一定规模的数据中心的数据备份要涉及到多种UNIX平台和不同的数据库类型，可以想象每天的备份工作对于管理员来说都是一个挑战。

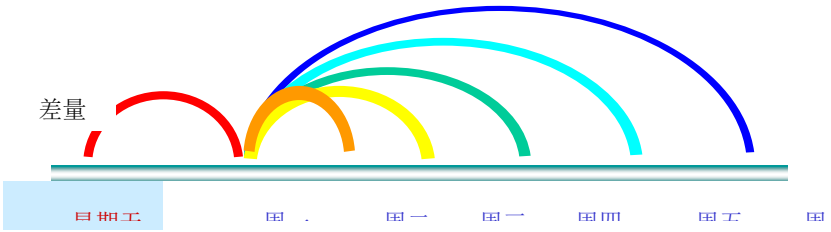
备份策略制定是备份工作的重要部分。一般来说需要备份的数据存在一个2/8原则，即20%的数据被更新的概率是80%。这个原则告诉我们，每次备份都完整的复制所有数据是一种非常不合理的做法。事实上，真实环境中的备份工作往往是基于一次完全备份之后的增量或差量备份。

完全备份很好理解，即把所有数据进行一次完整的备份，当进行恢复的时候只需要一盘磁带；

增量备份是只有那些在上次完全备份或者增量备份后被修改了的文件才会被备份，如下图，优点是备份数据量小，需要的时间短，缺点是恢复的时候需要多盘磁带，出问题的风险较大，



差量备份是备份那些自从上次完全备份之后被修改过的文件，如下图，因此从差量备份中恢复速度是很快的，因为只需要两份磁带——最后一次完全备份和最后一次差量备份，缺点是每次备份需要的时间较长。



备份窗口是在进行备份操作时，应用系统可以接受的最长备份时间，对于某些5X8类型的非关键应用备份窗口可以很大，但是对于7X24小时的应用备份窗口就会很小。

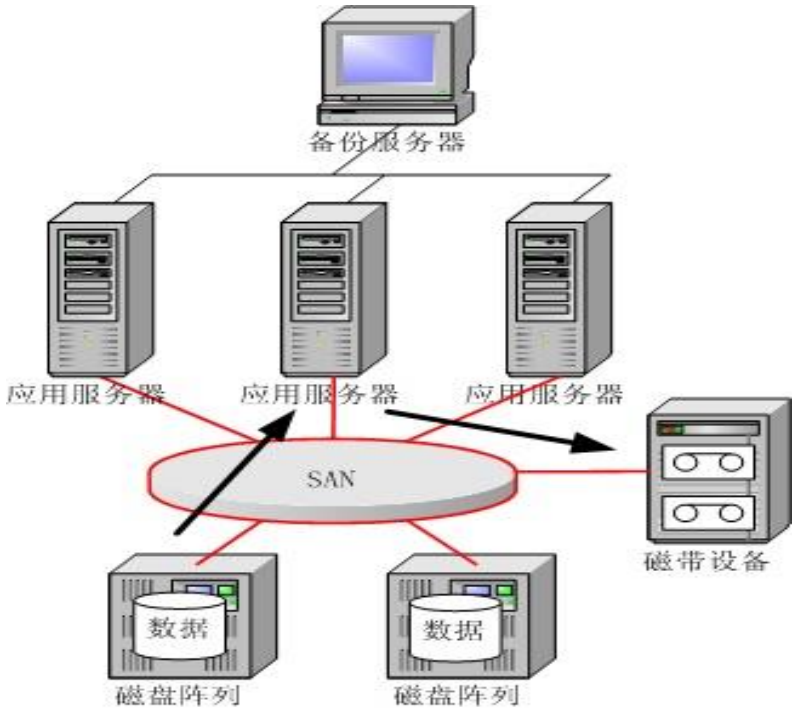
8.4 LAN Free 和 Serverless 备份

所谓LAN Free Backup顾名思义，就是指释放网络资源的数据备份方式。在SAN架构中，LAN Free Backup的实现机制一般如下图1所示。

备份服务器向应用服务器发送指令和信息，指挥应用服务器将数据直接从磁盘阵列中备份到磁带库中。在这个过程中，庞大的备份数据流没有流经网络，为网络节约了宝贵的带宽资源。在NAS架构中，情形十分类似，磁带库直接连接在NAS文件服务器上，备份服务器通过NDMP协议，指挥NAS文件服务器将数据备份到磁带库中。细心观察之下会发现，这两种方式虽然都节约了网络资源，但却增加了服务器的工作负荷，缺点是价格非常昂贵，大多数备份软件的LAN Free功能选项都需要用户付出高昂的价格。

Serverless Backup技术是以全面的释放网络和服务器资源为目的的，技术核心就是在SAN的交换层实现数据的复制工作，这样备份数据不仅无需经过网络，而且也不必经过应用服务器的总线，完全的保证了网络和应用服务器的高效运行。但是现实情况却没有这么理想，Serverless Backup技术目前只能停留在纸面上，实际实施效果很差，完全不需要主机干预还不现实。

图1 LAN Free Backup典型组网图



8.5 主流备份软件和介质

备份软件厂商中头把交椅当属Veritas公司。这家公司经过近几年的发展和并购，在备份软件市场已经占据了四成左右的份额。其备份产品主要是两个系列——高端的NetBackup和低端的Backup Exec。其中NetBackup适用于中型和大型的存储系统，可以广泛的支持各种开放平台。NetBackup还支持复杂的网络备份方式和LAN Free的数据备份，其技术先进性是业界共同认可的。

Backup Exec是原Seagate Soft公司的产品，在Windows平台具有相当的普及率和认可度，微软公司不仅在公司内部全面采用这款产品进行数据保护，还将其简化版打包在Windows操作系统中，我们现在在Windows系统中使用的“备份”功能，就是OEM自Backup Exec的简化版。2000年初，Veritas收购了Seagate Soft之后，在原来的基础上对这个产品进一步丰富和加强，现在，这款产品在低端市场的占用率已经稳稳的占据第一的位置。

Legato公司是备份领域内仅次于Veritas公司的主要厂商。作为专业的备份软件厂商，Legato公司拥有着比Veritas公司更久的历史，这使其具有了相当的竞争优势，一些大型应用的产品中涉及到备份的部分都会率先考虑与Legato的接口问题。而且，像Oracle等一些数据

库应用干脆内置集成了Legato公司的备份引擎。这些因素使得Legato公司成为了高端备份软件领域中的一面旗帜。在高端市场这一领域，Legato公司与Veritas公司一样具有极强的技术和市场实力，两家公司在高端市场的争夺一直难分伯仲。

Legato公司的备份软件产品以NetWorker系列为主线，与NetBackup一样，NetWorker也是适用于大型的复杂网络环境，具有各种先进的备份技术机制，广泛的支持各种开放系统平台。值得一提的是， NetWorker中的Cellestra技术第一个在产品上实现了Serverless Backup的思想。仅就备份技术的先进性而言，Legato公司是有实力可以挑战任何强大对手的。

除了Veritas和Legato这备份领域的两大巨头之外，IBM Tivoli也是重要角色之一。其Tivoli Storage Manager产品是高端备份产品中的有力竞争者。与Veritas的NetBackup和Legato的NetWorker相比，Tivoli Storage Manager更多的适用于IBM主机为主的系统平台，但其强大的网络备份功能觉对可以胜任任何大规模的海量存储系统的备份需要。

CA公司是软件领域的一个巨无霸企业，虽然主要精力没有放在存储技术方面，但其原来的备份软件ARCServe仍然在低端市场具有相当广泛的影响力。近年来，随着存储市场的发展，CA公司重新调整策略，并购了一些备份软件厂商，整合之后今年推出了新一代备份产品——BrightStor，这款产品的定位直指中高端市场，看来CA公司誓要在高端市场与Veritas和Legato一决雌雄。

表1 主流磁带驱动技术指标

	单盘容量（GB）	持续传输率（MB/s）	记录方式	介质类型	介质寿命（年）
LTO	100	15	线形	MP	30
SuperDLT	110	11	线形	AMP	>30
9940	60	10	线形	AMP	15-30
3590	20/40/60	14	线形	AMP	15-30
AIT-3	100	12	螺旋扫描	AME	>30
DTF-2	200/60	24	螺旋扫描	AMP	>30
Mammoth-2	60	12	螺旋扫描	AME	30

带机、带库厂商及产品

备份设备的生产厂家很多，每个厂家都有着较长的产品线，由于篇幅所限，我们不可能一一列举。这里主要介绍那些国际知名的、国内有影响力的带机和带库原厂商及其主打产品。

目前，带机正在朝快的数据传输速度和高的单盘磁带存储容量方向发展，具有主流驱动技术的带机厂商包括Quantum、Exabyte和Sony等。

Quantum带机在中档产品中占据了市场大部分份额，但其中很大一部分走了OEM的销售渠道。其自动加载机SuperLoader可将多个备份目标集中到一个共享的自动系统中，降低处理成本，而基于磁盘（备份介质是磁盘）又具有磁带海量特性的近线备份设备DX30可显著缩短备份与恢复时间。

Exabyte 的磁带驱动技术包括 8mm Mammoth 和 VXA 技术，VXA 是定位低端的新的磁带技术，它以包的格式读写数据，并可对磁带上的数据记录区进行无空隙扫描，具有高质量、高可靠性、低成本等性能特点。其中 VXA－1 带机专为苹果机设计的存储方案；VXA－2 同样具有较高的性价比，并具有 12MB/s 传输速率及 160GB 容量，与 VXA－1 向下兼容。

这里我们有必要讲一讲 Sony 的基于 AIT 技术的带机产品：AIT－1、AIT－2 和 AIT－3，其中 AIT－3 是高性能和大容量的新存储方案，容量（未压缩）为 100GB，速率为 12MB/s，而且能够与 AIT－1、AIT－2 完全读和写逆向兼容，并具有分层磁头、创新性的磁带内存储器(MIC) 驱动器接口系统等多项专利技术，提高磁轨密度和存储速度。

磁带库厂商相对品牌较多，用户的选择空间也更大一些。目前主流的磁带库厂商主要有STK，Quantum，Exabyte和IBM等。

在带库厂商中，市场份额最大的当属美国存储技术公司（StorageTek，STK）。STK目前最主要的产品线是L系列，包括L20、L40、L80、L180、L700、L5500，从最小20磁带槽位到最大5500磁带槽位。在其入门级产品上，支持LTO、DLT和SuperDLT等开放技术，只有在高端产品上才同时支持其自身拥有的9840、9940驱动技术。

Quantum拥有DLT、SuperDLT技术，其用户基础和发展前景都很好。其P系列的主打产品P4000和P7000分别可以支持几百槽位和十几个驱动器，适合于企业级用户；M系列是模块化的产品，可根据用户系统需求的增长灵活扩展带库的容量和性能，M1500可从20槽位扩展到200槽位，M2500则可从100槽位扩展到300槽位，非常适合于那些快速发展的中小企业。美中不足的是，ATL对超大容量的解决方案不是非常理想，在这一部分市场上的竞争力较弱。

8mm是安百特（Exabyte）公司的独立技术，具有速度快、容量大、可靠性高、价廉、体积小等特点，主要用于带库，其8mm带库的智能机械臂系统可任意存取磁带，采用模块化设计，产品线全，从VXA自动化/驱动器产品系列AutoPak230/115/110、VXA-1/1到Mammoth

Tape自动化/驱动器产品系列X200/80/430M/215M/EZ17、M2/Mammoth/Eliant 820，容量从单盘（非压缩）33GB到整库12TB，涵盖由低到高的用户市场，可实现无人值守自动数据存储管理，适用于服务器备份、网络备份、自动归档、分级存储管理及图形图像等领域。

IBM，众所周知，生产和销售所有IT类产品，当然也包括带库产品。IBM的带库和带机产品大体可分2个系列：用于IBM环境的和用于开放环境的。如IBM的3494、3575等带库只支持其专用的驱动器，开放性差，虽然这些带库产品也支持HP、SUN等主流服务器平台，但实际上几乎只用在IBM环境中。随着SAN技术的普及，追求开放性和互联性成为存储行业的潮流。结合LTO驱动技术的投产，IBM为其开放存储系统解决方案推出了新的带库系列——3583和3584。表2列出了上述带库生产厂家部分产品的参数。

表2 主流带库产品参数表

厂家	产品型号	磁带槽位	最大驱动器数	驱动技术类型	是否可级连
STK	L20	10、20	2	DLT8000 SDLT LTO Ultrium 9840、9940	否
	L40	20、40	4		否
	L80	40、60、80	8		否
	L180	174	10		否
	L700	678、1344	24、40		是
	L5500	5500（单台）	80		是
Quantum	M1500	20-200（10模块）	20	DLT8000 SDLT LTO Ultrium	是
	M2500	100-300（3模块）	18		是
	P4000	171-3	10		是

		22（单台）			
	P7000	399-679（单台）	16		是
IBM	3583	72	6	LTO Ultrium	否
	3584	2481（6模块）	72		是
	3494	160-240	32	3490e、3590	是
Exabyte	X200 X80 430M 215M	200 80 30 15	10 8 4 2	Mammoth-2、Mammoth Mammoth-2、Mammoth Mammoth Mammoth－2	

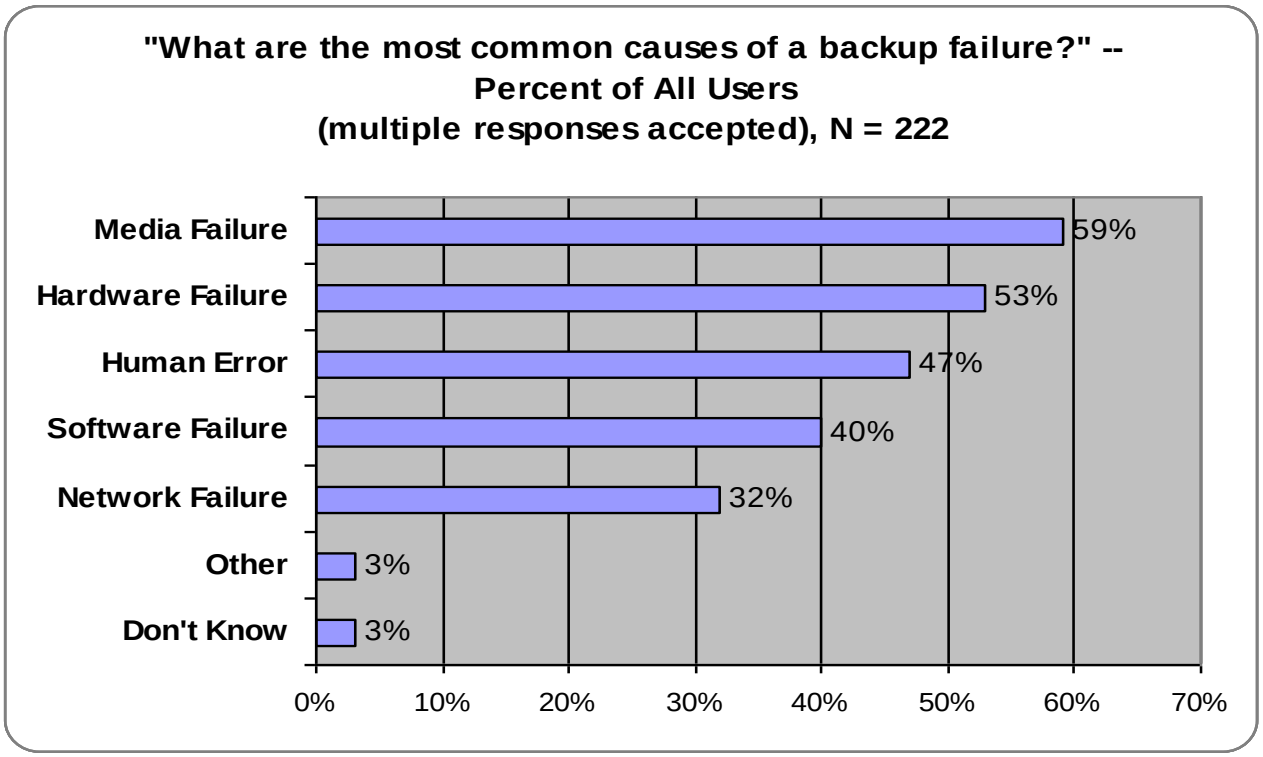
8.6 备份技术新趋势

D2D2T是Disk to Disk toTape的缩写，即数据备份从磁盘阵列到磁盘库到磁带的过程。传统的磁带

备份总是会带给用户以下苦恼：（图2）

- 备份速度慢，备份窗口冗长
- 备份的根本目的在于恢复，而磁带的恢复速度很慢，对于 TB 级的数据恢复等待时间过长
- 磁带介质受灰尘、温度、湿度影响很大，难以保证已经离线保存的磁带在需要的时候可以正常工作
- 磁带库的机械手等物理设备的故障率和磨损率相对电子元件较高

相信长期从事磁带备份工作的管理员（尤其是大数据量关键应用的磁带备份）对以上几点都会深有感触，尤其是当在线数据受到破坏，需要依靠磁带备份来恢复正常生产的时候，大家都会为能否顺利恢复数据捏一把汗。



有什么办法可以解决磁带备份固有的劣势呢？随着磁盘容量的增长价格的下降，使用磁盘备份作为磁带备份的补充甚至替代都成为可能（参考下图3），当然磁带体积小，便于归档等特点是磁盘设备不具备的，因此D2D2T即磁盘到磁盘到磁带备份方式有效地中和了磁盘备份和磁带备份的优点，在线数据保存在高速磁盘阵列，备份数据首先保存在性价比较高的SATA磁盘阵列中，然后定期将磁盘备份的数据保存到磁带上，这样既缩短了备份窗口又增强了数据恢复的可靠性，见图4。

图3 磁盘与磁带价格变化趋势

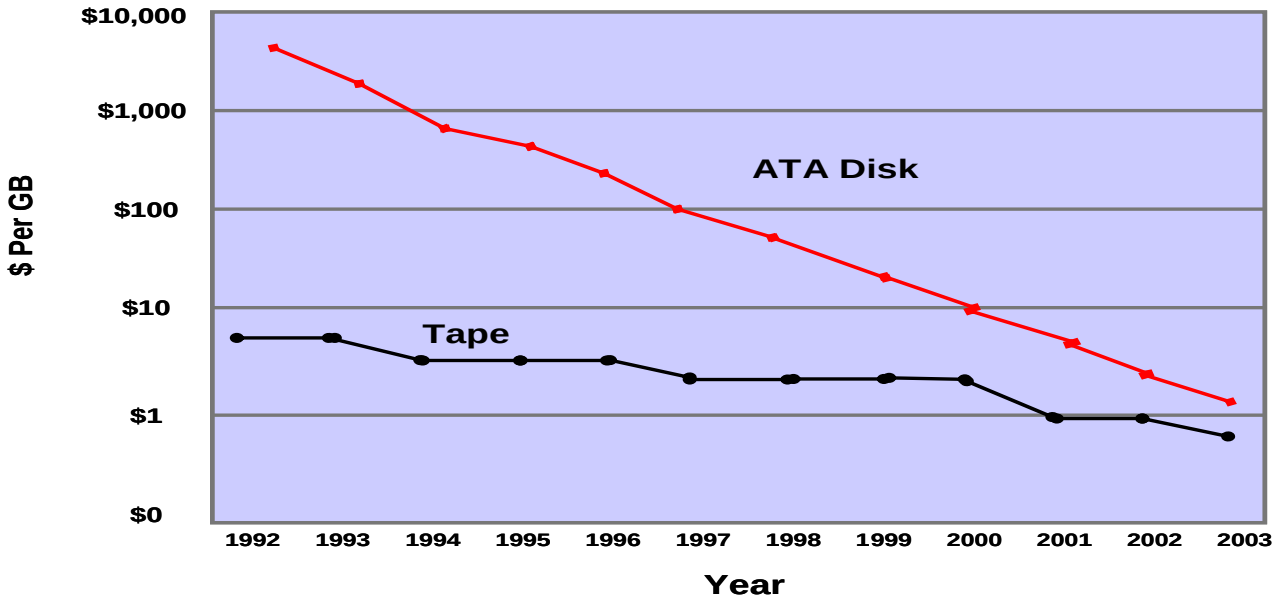
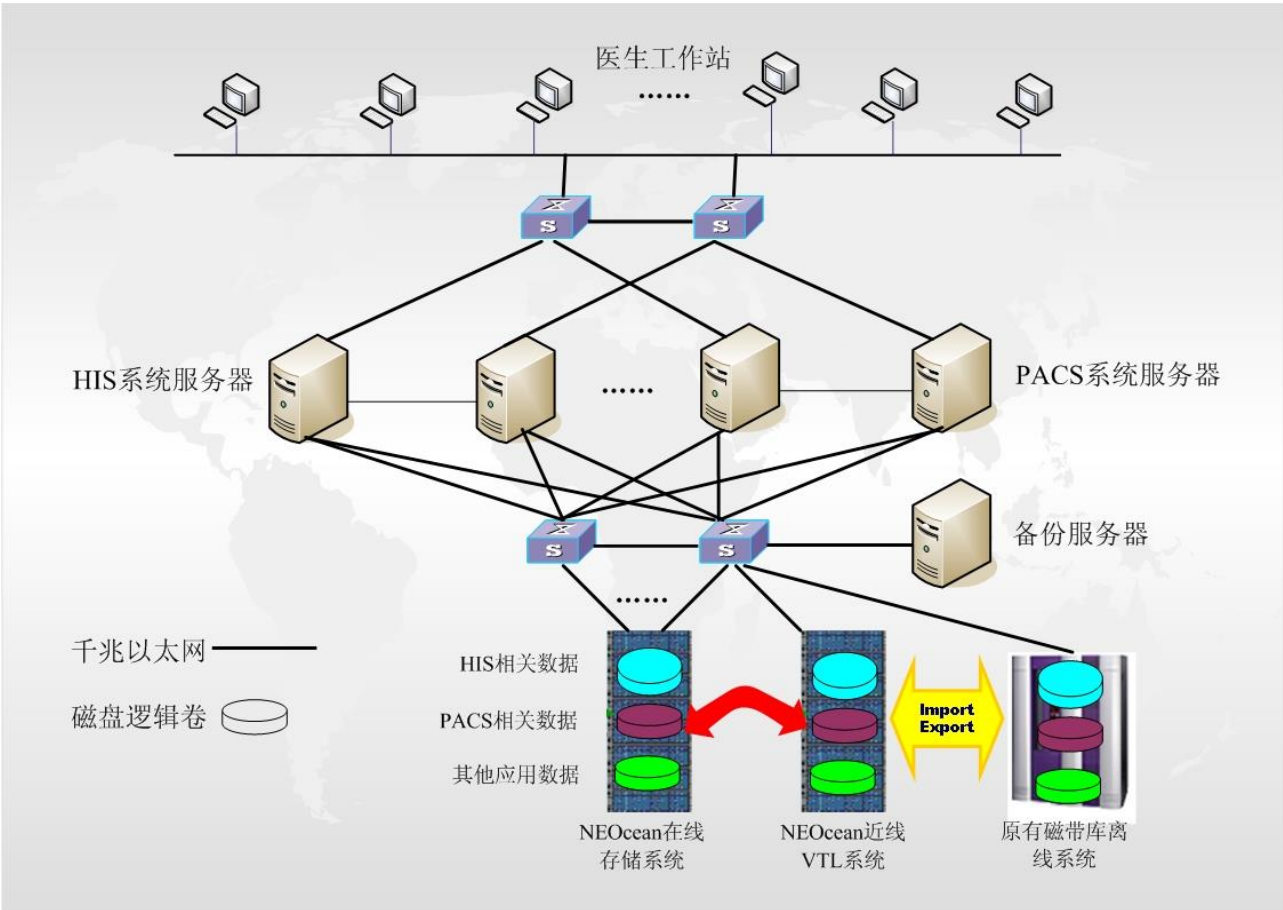


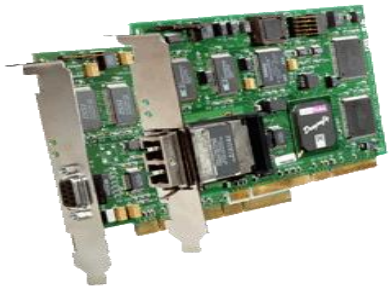
图4 典型D2D2T组网图



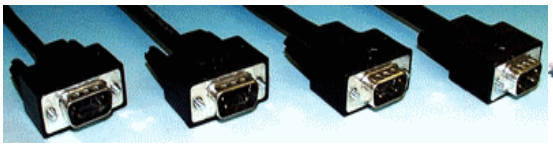
第9章 存储连接设备

9.1 HBA 卡介绍

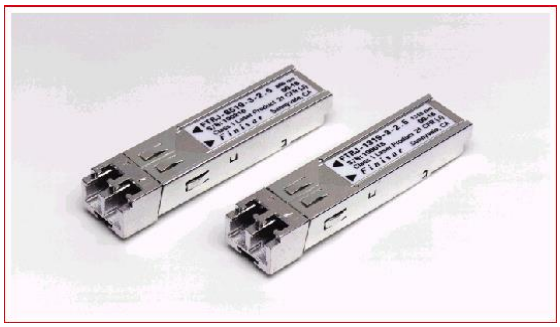
9.1.1 FC HBA 相关知识：



- 光纤信道主机总线适配器的缩写，主要用于主机与 FC 设备之间的连接；
- FC HBA 卡使用的传输介质可以分为光缆和铜缆两种；
- 光缆分为单模长波和多模短波两种，单模长波最大传输距离十公里，多模短波最大传输距离 500 米（直径 50 微米）、175 米（直径 62.5 微米）；
- 铜缆最大传输距离 30 米
- 光缆接口分为 SC、LC 两种，分别为 1Gb 和 2Gb 接口



- 铜缆接口分为 DB9 和 HSSDC 两种，其中 DB9 为 1Gb 接口，HSSDC 分为 1Gb 和 2Gb 两种；
- Fibre Channel 还包括三种独立于介质的接口：Gigabit Link Modules——将并行信号转换为串行信号，包括串行编码和串行解码功能；Gigabit Interface Converters——提供一个串行编码和串行解码的串行接口，可以热插拔，常见接口为 SC 和 DB9 两种；



Media Interface Adapters——将DB9铜口转换为光缆接口；

图片暂缺

9.1.2 主要 HBA 卡厂商

HBA卡主要厂商为Emulex、Qlogic、LSI、Adaptec、JNI、Agilent（安捷伦）等；

9.1.3 iSCSI HBA 相关知识：

- iSCSI 主机总线适配器的缩写，将硬件 iSCSI initiator 集成到板卡上，利用 TCP/IP 卸载引擎在适配卡上完成数据处理，作用在于减轻主机 CPU 负载，也提供了高可用环境下硬件多通道功能和服务器远程引导功能；
- 接口类型与以太网原有接口完全相同；
- 大多数 iSCSI HBA 卡都具有 TOE 功能；

TOE（TCP/IP Offload Engine）相关知识：

- TCP/IP 卸载引擎的缩写，TOE 网卡将 TCP/IP 处理脱离主机 CPU，在 TOE 网卡上由专用芯片完成 TCP/IP 处理和数据包创建，减轻主机 CPU 的负载；

表1 iSCSI HBA与TOE网卡的对比

	接口类型	硬件iSCSI initiator	TOE处理芯片
iSCSI HBA	RJ45、光口	有	有
TOE 网卡	RJ45、光口	没有	有

上表简单列举了TOE网卡和iSCSI HBA卡的区别，随着iSCSI产品的普及，目前大多数

iSCSI HBA也同时具备TOE功能，把二者的优点整合在一起。

9.1.4 iSCSI HBA 和 TOE 网卡主要厂商

TOE网卡以Intel、Alacritec、Qlogic、Adaptec

9.2 FC 连接设备介绍

9.2.1 FC HUB 相关知识：

- 类似以太网 HUB，内部为仲裁环拓扑（Arbitrated Loop），所有连接到 FC HUB 上的设备共享全部总线带宽，由于自身协议的原因，最多支持 127 个地址；

9.2.2 FC Switch 相关知识：

- FC 交换机，内部为 Fabric 拓扑，每端口独占带宽，理论上可以连接 1600 万个设备。

9.2.3 FC Director 相关知识：

- 多端口、高带宽网络的 Switch，用于提供最高的可用性，一般用于 FC 骨干网中多个 SAN 之间的连接。Director 中的每个部件的失灵不会影响正常应用，具有全冗余、热插拔部件，能将宕机时间最小化。

9.2.4 iSCSI-FC 存储路由器

- 结合 IP 和 FC 交换技术的存储网络平台，配置以太网端口和 FC 通信端口，通过逻辑单元号（LUN）映射实现 iSCSI 和 FC 之间的协议转换。

9.2.5 FC Switch 和 FC Director 主要厂商

McData、Brocade、Cisco等；

第10章 信息生命周期

10.1 什么是信息生命周期

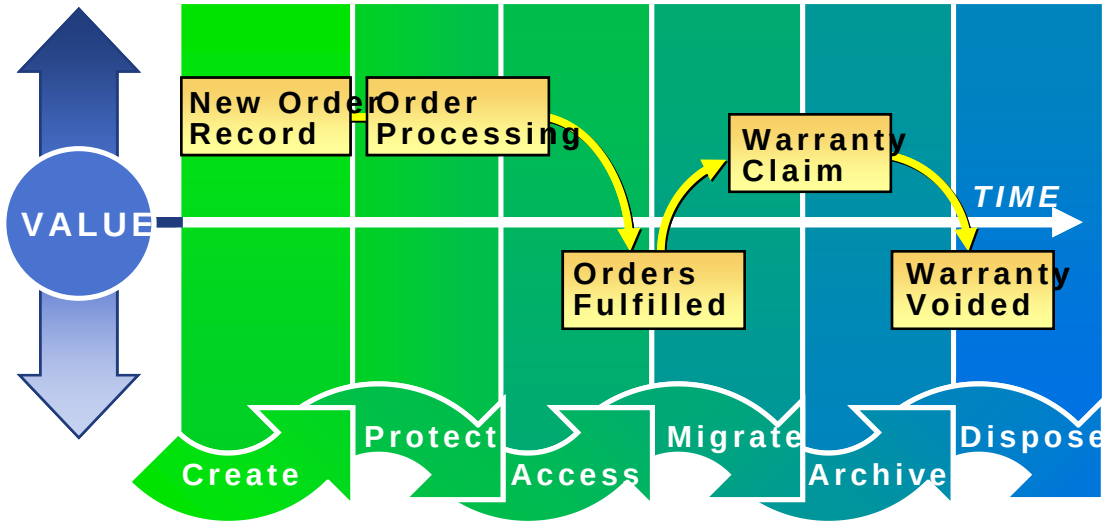
信息生命周期是STK、EMC等公司于2003年提出的概念，不同厂商对信息生命周期的定义都会有相同，但是其目的是统一的：适应信息在其生命周期不同阶段内价值的变化，利用网络存储技术将数据自动保存于适当的介质中，实现资源优化、投资保护、法规遵从等要求，直接或简介为适应业务和应用的变化提供保障。

自然界的一切事物不论是具体的动物、植物、组织还是抽象的思想、数据、信息都要经历一个从出生到成熟直到消亡的过程，在不同的过程中事务自身的很多特点都在发生变化。信息生命周期关注的焦点自然是数据，一般来讲数据在自身的生命周期内需要经过六个阶段：

- 数据生成阶段
- 数据保护阶段（随应用重要性不同，采取不同级别的保护措施）
- 数据访问阶段（随着访问频率的不同，又可分为在线、近线和离线访问方式）
- 数据迁移阶段（随着信息重要性的逐渐衰减，将数据转移到不同的介质上保存，优

化资源分配，降低使用成本）

- 数据归档阶段（将积累的数据归档保存，为今后的决策分析、数据挖掘提供素材，同时也要考虑逐渐健全的法规遵从要求）
- 数据销毁阶段（在法规允许的范围内，销毁不再有价值的数据，释放资源）



10.2 信息生命周期的实现

一切数字化的信息最终都要保存于存储系统中，因此信息生命周期管理的实现与存储系统的关系密不可分，ILM的实现需要经过三个阶段：

- 统一存储资源管理：首先实现存储系统的整合，消除 Internal 存储和 DAS 存储方式，解决异构存储系统的互操作和统一管理问题；
- 针对不同应用系统分别实现 ILM：根据不同应用系统的特点实现该系统所属数据的分级存储，在存储资源统一管理的前提下实现数据的自动分级存储（在线、近线、离线、归档）；
- 在整个组织范围内实现 ILM：在第二步基础上，以整个应用系统为核心，实现存储系统资源化，自动适应应用变化的需求，降低管理成本，保护原有投资；

10.3 实现 ILM 的技术保障和面临的挑战

克服异构存储系统之间的兼容性问题、统一管理界面问题、互操作问题等是目前ILM实施面临的最大挑战，目前存储行业的顶尖厂商都在向这个方向发展，如HDS的通用存储平台、IBM的DS8000、EMC公司倡导的SRM软件和SMI－S标准等，但是如果各大厂商之间不能统一标准，我们就将看到基于不同标准下的***通用存储平台，***后果只能是

在存储平台这个更高的层次上形成新一轮异构怪圈。

10.4 信息生命周期管理现状

以北美、日本和西欧为代表的高度信息化地区，其大型组织和公司的信息系统规模庞大，存储系统异构化问题非常普遍，在同一个数据中心中大量异构存储产品共存，这些产品厂商不同、型号不同、年代不同，性能差异、运行状态差异、接口类型差异等问题困扰着IT管理人员，由此带来的问题和隐患严重制约着业务和应用系统的发展，因此开放的、统一的存储系统平台显得尤为重要，目前以EMC、IBM、HDS、HP、STK等公司为主在大力推广ILM的概念和应用；

在国内，存储系统的建设还处于起步阶段，真正有信息生命周期管理需求的用户还属于稀有资源，在很多企业和组织的信息中心，我们经常可以看到双机＋DAS磁盘阵列或者简单的SAN系统部署，他们关注的问题更多是现有的存储系统是否能够适应应用未来几年内的变化。因此ILM作为IT系统发展的大趋势值得我们去关注和学习，但是我们同时应该认识到距离实现理想中的信息生命周期管理，还有很长的路要走。

10.5 法规遵从与信息生命周期管理

对积累的大量历史信息进行分析，可以得到有助于企业未来发展的决策已经是大家普遍接受的事实。但是对于金融机构、医疗行业、跨国公司等组织来说，他们的信息如交易记录、电子病例、电子邮件等信息需要通过法律的形式来规定其保存形式、保存年限、保存地点等要求。信息在其生命周期中的“晚年”并不是象从前那样被逐渐替代和销毁，而是处于一种长期休眠的状态，而且现在的信息保留法规均以数据的业务价值和法律价值为基础，而不仅仅是信息被引用的概率了，与信息生命周期管理相关的法规主要有HIPPA、21 CFR Part 11、DISC PD 0008：1999、SEC 17a－4等。

10.6 与信息生命周期相关的存储技术

10.6.1 固定内容管理：

在过去30年中信息技术的反展经历了三个重要阶段：上世纪70年代以大型机和服务器为核心的阶段、上世纪80年代以个人计算机发展为主的阶段、上世纪90年代后期以互联网为核心的浪潮，现在我们正面临着一个新浪潮的转折点——以内容为核心的信息技术革命。

至2004年底全世界信息量已经接近600亿GB，而且信息量增长几乎是以每年成倍的速度在增长，在如此快速增长的信息中，固定内容占有75%左右的比例，例如文档、邮件、图片、影像资料等等，而且超过90%以上的新增信息将会以数字化的形式保存。

固定内容信息的爆炸式增长催生了固定内容管理技术的发展，例如EMC公司的Centera、IBM公司的Content Manager，国内厂商则以TRS和方正博思为代表。

10.6.2 WORM：

WORM是Write Once Read Many的缩写，即一次写入多次读取，WORM技术是实现固定内容管理的一种技术手段，对于交易记录、电子病例、医学影像文件、电子邮件等类型的文件其内容可能用于信用评估、调查证据、裁决纠纷等行为的证据，因此要求从存储设备层次上杜绝对文件内容的更改，WORM技术允许管理员对保存在存储设备中的文件设定策略，例如设定在多长的时间内保存的文件内容不得被修改和删除，WORM技术尤其适用于实现法规遵从对存储系统的要求。

10.7 怎样看待信息生命周期管理

通过上文的分析，我们可以看到，考虑到投入产出比于企业承受能力，信息生命周期管理目前尤其适合于存储设备异构化严重，导致管理维护成本据高不下，甚至影响到业务和应用系统发展规划的大型组织和机构，对于大多数存储系统建设仍处于起步阶段的用户来说，在系统规划阶段，应首要考虑建设一套适应变化能力强、扩展能力强、管理维护简便的存储系统，这是实现信息生命周期管理的基础，信息生命周期管理作为IT系统发展的大方向，需要我们在IT系统建设中付出长期的努力才能逐步实现。

第11章 其他存储技术及标准

11.1 SMI－S

SMI-S （Storage Management Interface Specification 存储管理接口标准）是SNIA开发的一种标准管理接口，旨在减轻多厂商SAN（存储区域网络）环境的管理负担。SMI-S 为各种网络组件提供了一个通用管理接口，减小了SAN管理的复杂性。SMI-S发布至今已经取得多家SNIA成员企业的认可与支持。它将为网络存储行业定义一个全新且开放的开发模式，为其各组成部分丰富管理功能、提高互操作性，帮助各成员提交可管理、可互操作的存储网络解决方案。

SMI-S的目标是在存储网络中的存储设备和管理软件之间提供标准化的通信方式，从而使存储管理实现厂商无关性(vendor-neutral)，提高管理效率、降低管理成本，促进存储网络的发展。

- 对行业用户的意义所在：SMI-S使用户能够在SAN中轻松集成和管理来自多厂商的组件，从而提升了灵活性、可管理性和可靠性；同时，用户的资源利用率也将获得极大的提高。
- 对厂商的意义所在：广泛采用SMI-S将减少了选择和供应产品的复杂性，同时为基于政策的管理奠定了坚实的基础。厂商能够全神贯注于附加值功能上，而不必为异构和专有接口开发整合所需的技术支持。SMI-S在统一理解存储管理上对所有厂商都是至关重要的。有了一个公共平台，厂商就可以加速产品的开发进程，并且终端用户可以更自由地选择厂商，同时也降低了复杂性。

此外，SMI-S还制定了其他一些应用模块，它们除了可以使开发和测试过程更加简化以外，还能用于管理存储网络，为存储和软件工业提供新的发展思路。简言之，这些包括：管理软件的共存， 通过全新的系统架构，让多种管理软件在同一存储网络下共存；多层资源管理，大型存储网络中， 同样的服务经常在多层有所提供；基于策略的管理；互连的独立性；无缝集成；集成了安全性的管理；灵活的管理授权机制等等。

11.2 CDP（持续数据保护）

CDP是一个实时的数据备份系统，它自动地将应用数据的所有变化实时的纪录下来。它的关键是将每一个应用数据的变化加上了时间索引。这样，当出现数据丢失、数据损坏、或者安全问题时，就可能恢复到最近的完好数据。CDP技术不断在进步，在数据损耗的情况下，其允许快速的数据恢复，并实现把数据还原到生命周期任一指定点。

目前从事CDP产品的厂商规模较小，Veritas等大厂商也在密切关注这一领域，很可能通过收购的方式补充为自己的产品。

11.3 虚拟存储

虚拟存储就是整合各种存储物理设备为一整体，提供永久保存数据并提供能被用户调用的功能，即在公共控制平台下存储设备的一个集合体。虚拟存储分为如下四种：基于主机或服务器的虚拟化、基于磁盘或磁盘子系统的虚拟化、基于网络的虚拟化和基于交换机或路由器的虚拟技术。

存储技术发展到今天，随着数据的不断增加和存储网络化、集中化的发展，面临越来越多的问题，其中最为重要的一点就是，面对复杂的异构IT平台，如何对不同存储设备间的资源进行共享，节约资源，使存储管理简单化？最重要的是，现行存储厂商开发的虚拟化产品是否真的可以做到，用户采用虚拟化技术所希望的那样，能在一个异构的存储环境里，支持所有或大多数的存储硬件，这样用户不必花费更多的开支去替换原有的设备，这才是用户关注青睐存储虚拟化产品的最终目的。

在虚拟存储环境下，存储将不依赖于地理位置和互连体系结构，变得可视、具有弹性、可移动而且安全。用户将不必担心使用的是何种设备，他们所需要关心的将仅仅是使用空间的大小，而这也仅仅是管理存储空间的分配，而不是具体的某一物理设备。

目前倡导虚拟存储的厂商众多，传统厂商以HDS、HP、SUN为代表，纯软件厂商以FalconStor，StoreAge，Datacore等为代表。

11.4 网格计算

网格计算是一个集成了多站点，支持应用、支持协作的分布式计算技术，网格实际上是一个大型的分布计算架构，它集成了多种标准资源，多个域环境，它们可能是位于同一地点，但更多情况是位于不同地点的，它支持多种应用用来实现资源共享，以及促进协作；网格关系到应用程序、商业逻辑的架构，企业的业务需要经过一段时间进行转换，才能适应网格架构。这对于计算技术来说，实际上是一次质的跨越。在网格成熟的时期，我们不会再关心CPU或者其它硬件的性能，甚至也不再讨论应用程序的功能和效率了，唯一要关心的就是服务。

11.5 高性能计算

简单的说，高性能计算(High-Performance Computing)是计算机科学的一个分支，它致力于开发超级计算机，研究并行算法和开发相关软件。高性能计算主要研究如下两类问题：

- 大规模科学问题，象天气预报、地形分析和生物制药等；
- 存储和处理海量数据，象数据挖掘、图象处理和基因测序；

高性能计算应用分类：高性能计算从数据I/O类型可分为I/O集中式计算和I/O分布式计算。

- I/O 集中式计算是把一个任务分成若干可以并行执行的子任务，各个子任务间在计算过程中较少进行数据交换，此类型应用 I/O 瓶颈在节点与存储系统之间，要求存储系统提供尽可能高的吞吐量和 I/O 处理能力，所以把这类计算称为 I/O 集中式计算。

I/O分布式计算的子任务间联系很紧密，需要频繁进行节点间数据交换，由于节点间通常用以太网方式通讯，因此此类型应用I/O瓶颈往往在内部以太网。

11.6 负载均衡

随着应用系统的不断发展，服务器所承受的压力也在不断的加大。通过采用服务器集群技术，将客户端的访问请求在多个应用服务器之间进行负载分担，有效的缓解了单个服务器

的压力。应用服务器跟客户端之间的访问带宽增加了，就对应用服务器跟存储系统之间的访问带宽提出了更高的要求。如何解决这一问题呢？采用基于MPIO的负载均衡技术是一个不错的选择。

MPIO就是在应用主机和存储设备之间建立多条数据通路，以提高数据访问的带宽。具体做法上，首先需要在应用主机配置多个网络传输设备（网卡、HBA等），建立多条物理链路。然后安装必要的MPIO软件，将主机应用在多条物理链路上连接到相关的数据卷。实际使用中，应用端将数据通过多条通路发送到存储系统，在存储系统端再将多条通道上的数据汇总组合。通过这种方式，可以在同一时间通过多条通道对数据卷进行访问，有效的增加了带宽，实现了负载均衡功能。

第12章 常见主机及操作系统

12.1 主机架构及操作系统概述

12.1.1 主机架构

一般来说. 主机按照系统架构来分. 主要分为两种类型. 一种是基于X86体系系统架构. 采用AMD的Opteron和Intel的Xeon/ Itanium等处理芯片，支持标准的Windows和Linux操作系统的. 是一个通用开放的系统.另一种是基于专有的64位RISC芯片体系系统架构. 由各厂商自己开发芯片和操作系统，各家产品互不兼容，也不兼容大量x86平台上的软件，一般采用UNIX操作系统.一般都称为小型机.像美国Sun、日本Fujitsu等公司的小型机是基于SPARC处理器架构,而美国HP公司的则是基于PA－RISC架构,Compaq公司是Alpha架构,IBM和SGI等的也都各不相同;其中I／O总线也不相同,Fujitsu是PCI,Sun是SBUS,等等, 这就意味着各公司小型机机器上的插卡,如网卡、显示卡、SCSI卡等可能也是专用的。

12.1.2 操作系统

操作系统是计算机系统中的管理软件，负责管理计算机系统中的所有资源.主要用于管理主机的进程管理与协作、内存管理、文件系统与输入输出、系统保护与安全等一般来说. 常见主机系统主要分为两大系列. 一类是基于Microsoft的Windows系列. 包括Windows 2000/2003/NT/XP等. 主要是用于基于Intel处理器系统架构的PC服务器.一类是基于Unix.

依据不同的应用环境和不同的主机架构. 分为TurboLinux 、RedHat、BSD、 Solaris、SCO、HP-UX、 AIX、 Tru64、 IRIX、 MacOsX等等. 有些操作系统是专用于使用专用处理器架构的主机厂商. 例如Sun Solaris是基于使用了SPARC处理器架构的主机使用. Sun公司和Fujitsu公司的主机, HP－Unix是用于HP公司的主机, IBM是AIX, SGI公司的是IRIX等等,另外,PC服务器还可以支持现在流行的Linux、SCO Unix、Solaris For X86等Unix操作系统。

12.1.3 操作系统比较

- UNIX系列：是世界上应用最为广泛的一种多用户多任务操作系统。主要特点：功能强大、文件系统简练、可扩充性与可移植性好、网络通信功能强(TCP/IP)；使用UNIX用户一般是看中Unix操作系统的安全性、可靠性和专用服务器的高速运算能力；
- Windows系列：安全性、可靠性稍差. 但在操作界面的易用性、通用性、易开发性、大量的应用软件 支持上好；

12.2 常见主机厂商及常见产品介绍

12.2.1 IBM：

- 基于X86体系系统架构的IBM eServer x系列服务器：

处理器	CPU个数（最大）	内存（最大）	支持操作系统
Intel Itanium 处理器 Intel®Xeon™处理器 AMD Opteron处理器	16个	56GB	Microsoft Windows Server 2003、Windows 2000 Server/Advanced Server、 Novell、Netware、 Red Hat Linux、 SUSE LINUX.

◆ 基于专有的64位RISC芯片体系系统架构的IBM eServer® p系列/I系列/OpenPower系列服务器：

处理器	CPU个数（最大）	内存（最大）	支持操作系统
POWER4/POWER5™ 处理器	32个	512GB	Linux操作系统、AIX操作系统、OS400

12.2.2 SUN：

◆ 基于专有的64位RISC芯片体系系统架构的SUN Fire V系列/E系列/ PRIMEPOWER系列/Netra系列服务器：

处理器	CPU个数（最大）	内存（最大）	支持操作系统
UltraSPARC III/IV 处理器	128个	512GB	Solaris

◆ 基于X86体系系统架构的V20/V40服务器

处理器	CPU个数（最大）	内存	支持操作系统
AMD Opteron处理器	4个	32GB	Solaris Operating System, 32-bit and 64-bit Red Hat Linux, 64-bit SUSE Linux, Microsoft Windows, 和 VMware

12.2.3 Fujitsu:

◆ 基于专有的 64 位 RISC 芯片体系系统架构的 PRIMEPOWER 系列服务器：

处理器	CPU个数（最大）	内存（最大）	支持操作系统
SPARC64TM V 处理器	128个	512GB	Solaris

◆ 基于 X86 体系系统架构的 PRIMEQUEST 系列服务器

处理器	CPU个数（最大）	内存	支持操作系统
Intel Itanium 处理器	32个	512GB	Microsoft Windows Server 2003、Windows 2000 Server/Advanced Server、 Novell、 Netware、 Red Hat Linux、 SUSE LINUX.

12.2.4 HP:

◆ 基于专有的 64 位 RISC 芯片体系系统架构的 HP 9000 Superdome、AlphaServer GS、ES、DS 系列服务器：

处理器	CPU个数（最大）	内存（最大）	支持操作系统
PA-8800处理器	128个	1 TB	HP-UX 11i v1
EV7 Alpha处理器	64个	512GB	Tru64 UNIX,

			OpenVMS

◆ 基于 X86 体系系统架构的 ProLiant、Integrity 系列服务器

处理器	CPU个数（最大）	内存	支持操作系统
英特尔® 奔腾®4 处理器 Intel®Xeon™处理器 Intel Itanium 处理器	128个	1TB	Microsoft Windows NT 4.0 Microsoft Windows 2000 Server Microsoft .NET Server 和Web Server Red Hat Advanced Server v2.1 Red Hat 8.0 SuSE SLES 8 HP-UX 11i（Integrity系列）

12.3 操作系统应用特点

文件格式/操作系统	支持最大的分区容量
NTFS（Windows）	支持最大分区2TB，最大文件2TB
FAT16（Windows）	支持最大分区2GB，最大文件2GB
FAT32（Windows）	支持最大分区128GB，最大文件4GB
HPFS（OS/2）	支持最大分区2TB，最大文件2GB
EXT2和EXT3（Linux）	支持最大分区4TB，最大文件2GB
JFS（AIX）	支持最大分区4PB，最大文件4PB
XFS（IRIX）	9E（2的63次方）

第13章 常见数据库及应用系统

13.1 数据库厂商介绍

从支持信息管理到联机事务处理（OLTP），再到联机分析处理（OLAP）分类。数据库可分为小型数据库系统和大型数据库系统。其中小型数据库系统以早先的Foxpro, ACCESS, PARADOX等为代表，大型数据库系统以DB2, INGRES, ORACLE, INFORMIX, SYBASE, SQL SERVER等为代表。

13.1.1 Oracle

该公司成立于1977年，最初是一家专门开发数据库的公司。Oracle在数据库领域一直处于领先地位。1984年，首先将关系数据库转到了桌面计算机上。然后，Oracle的下一个版本，版本5，率先推出了分布式数据库、客户/服务器结构等崭新的概念。Oracle的版本6首创行锁定模式以及对称多处理计算机的支持。Oracle能在所有主要的平台（其中包括Windows）上运行，并且完全支持所有的工业标准.覆盖了大、中、小型机等几十种机型，可在VMS、DOS、UNIX、WINDOWS等多种操作系统下工作。Oracle数据库成为世界上使用最广泛的关系数据系统之一。

Oracle采用多线索多进程体系结构，直接在内核中支持分布式数据库操作、多线索处理、并行处理以及联机事务处理等。从结构上看，同时支持集中式多用户环境、Client/Server、分布式处理和Internet计算结构；从技术上说，为应用程序及系统开发人员实现了透明的网络环境、混合网络结构以及分布式数据管理等。

Oracle Database 10g 提供了四个版本，每个版本适用于不同的开发和部署环境。Oracle 还提供了额外的几种可选数据库产品，这些产品针对特殊的应用需求增强了 Oracle Database 10g 的功能。下面是 Oracle Database 10g 的可用版本：

Oracle数据库10g标准版1（Oracle Database 10g Standard Edition One ）为工作组、部门级和互联网/内联网应用程序提供了易用性和性能价格比。从针对小型商务的单服务器环境到大型的分布式部门环境，Oracle Database 10g Standard Edition 包含了构建关键商务的应用程序所必需的全部工具。Standard Edition One 仅许可在最高容量为两个处理器的服务器上使用。

Oracle数据库10g标准版（Oracle Database 10g Standard Edition ）提供了 Standard Edition One 的前的易用性、能力和性能，并且提供了对更大型的计算机和服务集群的支持。可以在最高容量为四个处理器的单台服务器上、或者在一个支持最多四个处理器的服务器的集群上使用。

Oracle数据库10g企业版（Oracle Database 10g Enterprise Edition ）为关键任务的应用程序（如大业务量的在线事务处理 (OLTP) 环境、查询密集的数据仓库和要求苛刻的互联网应用程序）提供了高效、可靠、安全的数据管理。Oracle 数据库企业版为企业提供了满足关键任务应用程序的可用性和可伸缩性需求的工具和功能。包含了 Oracle 数据库的所有组件，并且能够通过购买选项和程序包来进一步得到增强。

Oracle数据库10g个人版（Oracle Database 10g Personal Edition） 支持需要与Oracle数据库10g标准版1、Oracle 数据库标准版和 Oracle 数据库企业版完全兼容的单用户开发和部署。引入到个人工作站中，该数据库具有桌面产品通常具有的易用性和简单性。

特性汇总	标准版1	标准版	企业版或个人版

高可用性			
故障保护 配置和验证Windows集群，并通过与微软集群服务器集成的高可用性软件快速准确地自动恢复Oracle数据库和应用系统。	√	√	√
回闪查询 无需复杂、耗时的操作即可恢复更早版本的数据。	√	√	√
回闪表、数据库和事务查询 诊断和撤销错误操作，包括对单独一行所做的修改、由杂乱的事务导致的变化、对单个或多个表所做的修改（包括表的删除）以及对整个数据库所做的所有修改。			√
数据卫士 自动维护生产数据库的多个远程备份副本；恢复从生产环境到备份数据库的处理；极大地缩短了灾难情况下的宕机时间。			√
可伸缩性			
真正应用集群 跨多个相互连接或"集群的"服务器运行任意未做更改的打包或定制的应用系统。		√	**可选**
集成的集群件 利用一组通用、内置的集群服务创建和运行数据库集群。		√	√
自动工作负载管理 将服务连接请求发送给拥有最低负载的适当服务器；一旦发生故障，自动将幸存的服务器重新分配以用于服务。		√	√
Java和PL/SQL的本地编译 用Java和PL/SQL语言编写部署在数据库中的程序。	√	√	√
安全性			

密码管理 利用单一用户名和密码连接整个企业内的多个数据库。	✓	✓	✓
加密工具包 借助PL/SQL包加密和解密存储的数据。	✓	✓	✓
虚拟专用数据库 编写行级安全性程序；确保应用程序上下文的安全。			✓
细粒度审计 定义特定的审计策略，包括对错误数据的访问发出警告。			✓
应用软件开发			
Java支持 更快地执行Java应用程序，集成现有的软件资产，将Java/J2EE应用程序连接到支持网格的数据库，通过Web服务支持非连接的客户，并将本地数据与远程和动态数据结合起来。	✓	✓	✓
HTML DB 借助快速Web应用系统开发工具，开发和部署快速、安全的应用系统。	✓	✓	✓
全面的XML支持 通过对W3C XML数据模型提供支持，使存储和检索XML本地化；使用标准访问方法导航和查询XML。	✓	✓	✓
PL/SQL和JSP 利用服务器端Java和存储的程序语言；使用SQL确保安全、方便和无缝。	✓	✓	✓
COM自动化、微软事务服务器/COM+集成、ODBC和OLE DB 支持多种Windows数据访问方法。	✓	✓	✓
可管理性			
企业管理器 通过单一集成的控制台，基于Oracle产品系列管理和监控所有应用程序和系统。	✓	✓	✓

自动内存管理 自动管理Oracle数据库实例使用的共享内存。	√	√	√
自动存储管理 跨所有可用资源分配I/O负载，并通过垂直集成的文件系统和卷管理器优化性能；消除人工I/O调优。	√	√	√
自动撤消管理 监控所有Oracle系统的参数设置、安全设置、存储和文件空间条件的配置。	√	√	√
服务器管理的备份和恢复 借助Oracle恢复管理器（RMAN）简化、自动化并提高备份及恢复性能。	√	√	√
数据仓储			
数据压缩 在不影响查询时间的情况下压缩保存在关系表中的数据；降低磁盘系统成本。			√
Oracle分析函数 使用面向在线分析处理（OLAP）的内置分析工作空间。	√	√	√
可移动的表空间，包括跨平台 将一组表空间从一个数据库转移到另一个数据库，或者从一个数据库转移到自身的其他位置。			√
星形查询优化 加入一个事实表和大量维度表。	√	√	√
汇总管理--物化视图查询改写 当一个物化视图用来响应一个请求时，通过自动识别提高查询性能；透明地改写请求以使用物化视图。			√
集成			
Oracle流 在一个数据库内或从一个数据库到另一个数据库的数据流中实现数据、事务和事件的传播与管理。			√
高级队列 通过基于队列的发布－订阅功能，使数据库队	√	√	√

列能够充当持久的消息存储器。			
工作流 支持与完整的工作流管理系统基于业务流程的集成。	√	√	√
分布式查询/事务处理 在分布式数据库的两个或更多个不同节点上查询或更新数据。	√	√	√
内容管理			
超级搜索 跨多个信息库搜索和定位数据，包括： Oracle数据库、遵循ODBC的数据库、IMAP邮件服务器、HTML文档、磁盘文件等。	√	√	√
媒介物 开发、部署和管理包含具有最流行格式的富媒体内容的传统、Web和无线应用系统。	√	√	√
文本 构建文本查询应用系统和文档分类应用系统。	√	√	√
定位器 管理地理空间数据来利用地域商业价值。使用企业版的空间选项，你还能够支持最复杂的GIS部署。	√	√	√

安装Oracle数据库10g的软硬件需求如下：

适应环境	
硬件需求	最多支持 4 个 CPU 的单个服务器，或者一个集群中最多 4 个 CPU 的集群服务器
软件要求	操作系统：Linux，UNIX，Windows

13.1.2 DB2

是IBM公司的产品，起源于System R和System R*。他支持从PC到UNIX，从中小型机到大

型机；从IBM到非IBM（HP及SUN UNIX系统等）各种操作平台。他既可以在主机上以主/从方式独立运行，也可以在客户/服务器环境中运行。其中服务平台可以是OS/400，AIX，OS/2，HP-UNIX,SUN-Solaris等操作系统,客户机平台可以是OS/2或Windows，Dos，AIX，HP-UX，SUN Solaris等操作系统，既可以在主机上以主/从方式独立运行，也可以在客户/服务器环境中运行，当前的最新版本为DB2 Universal Database version 8.1。

在Windows平台上安装DB2 Universal Database v8.1的硬件需求如下：

硬件	需求
处理器	Pentium或兼容处理器
内存	RAM：最小256 MB 可能需要额外的内存支持
硬盘空间	典型安装：最少350 Mb 最小安装：最少100 MB 自定义安装：最小100 Mb 在簇较大的FAT驱动器上可能需要额外的磁盘空间

如果在UNIX系列操作系统上安装DB2 Universal Database v8.1，包括基于AIX的系统、HP-UX系统、Linux和Sun Solaris，硬件必须满足：

硬件	需求
处理器	AIX：IBM RISC/6000或者eServer pSeries HP-UX：HP 9000系列700或800系统 Linux：Intel 32位、64位，S/390 9672或更高，Multiprise 3000，eServer z-Series Solaris：基于Solaris UltraSPARC的计算机
内存	最小 256 MB RAM 可能需要更多的内存

硬盘空间	典型安装：最小450到550 MB 紧凑安装：最小350到400 MB 自定义安装：最小350到700 MB

DB2 Universal Database v8.1有下面六个版本

- DB2 Enterprise Server Edition (ESE)
- DB2 Workgroup Server Edition (WSE)
- DB2 Workgroup Server Unlimited Edition (WSUE)
- DB2 Personal Edition (PE)
- DB2 Universal Developer's Edition (UDE)
- DB2 Personal Developer's Edition (PDE)

它们的软件需求如下：

平台	操作系统版本	要求安装的补丁
Windows	Windows NT 4.0	Service Pack 6a或更高
Windows	Windows 2000	Windows Terminal Server需要Service Pack 2
Windows	Windows XP	—
AIX	AIX 4.3.3 (32-bit)	Maintenance Level 9或更高，以及APARs IY22308，Y32690，和IY33024
AIX	AIX 5L (32-bit)	Maintenance Level 2或更高
AIX	AIX 5.1.0 (32-bit)	Maintenance Level 2或更高，以及APARs IY31254，IY32217，IY32905，IY33023，和IY29345
AIX	AIX 5.1.0 (64-bit)	Maintenance Level 2或更高，以及APARs IY31254,IY32217,IY32905,Y33023,和IY32466
HP-UX	HP-UX 11i	December 2001 GOLDBASE11i, December 2001 GOLDAPPS11i bundles

Linux	kernel level 2.4.9或更高 Red Hat Linux 7.2 SuSE Linux SLES-7	—
Sun Solaris	Solaris 7（32-bit）	patch 106327-10
Sun Solaris	Solaris 7（64-bit）	patch 106300-11
Sun Solaris	Solaris 8（32-bit）	patch 108434-03 and 108528-12
Sun Solaris	Solaris 8（64-bit）	patch 108435-03 and 108528-12
Sun Solaris	Solaris 9	—

13.1.3 Sybase

Sybase公司是1986年成立的，属于数据库厂商的后起之秀,立足于在开放系统平台上研制具有Client/Server体系结构的数据库系统软件产品。当时的Sybase公司以满足联机事务处理(OLTP)应用的高性能要求为目标，同时加强联网对异构数据源的开放互联，使Sybase取得了很大成功，特别是Sybase有效的拓展市场行动，使Sybase曾一度成为世界数据库市场增长最快的产品。

Sybase采用单进程多线索体系结构，其核心是SQL Server（现在为与Microsoft SQL Server区别，改名为Adaptive Server Enterprise），在Client/ Server体系结构的基础上，架构了复制服务器(Replication Server)和多功能通用网关(Omni SQL Gateway，现在叫Enterprise CONNECT)等部件，从而形成了软件产品系列。

Sybase 公司的数据库Adaptive Server Enterprise是针对电子商务应用环境而推出的

企业智能型关系数据库管理系统，它具有开放的、可扩展的体系结构，易于使用的事务处理系统。目前最新的版本为Sybase Adaptive Server Enterprise 12.5（ASE 12.5），Sybase ASE 12.5支持传统OLTP繁重的数据库处理需要与决策支持系统的复杂数据操作需求，还提供了对Java存储过程的支持、支持Java用户自定义函数和数据类型，支持在服务器上通过Java访问远程服务器,可以实现在数据库中存储及搜索XML文档并实施Java应用，使客户可以在复杂的环境中实现事务密集型关键应用的智能管理。

安装Sybase Adaptive Server Enterprise 12.5（ASE 12.5）的软硬件需求如下：

适应环境	
硬件需求 处理器	Intel 32位、64位
软件要求	操作系统：NovellNetware,WindowsNT,UNIX,Linux，solaris x86,Unixware

13.1.4 MS SQL Server

SQL Server是Microsoft公司的一个关系数据库管理系统，从20 世纪80年代后期开始开发，最早起源于1987年的Sybase SQL Server,发布了用于Windows NT操作系统的SQL Server，将SQL Server移植到了Windows NT平台上。SQL Server只在Windows上运行，Microsoft这种专有策略的目标是将客户锁定到Windows环境中

目前最新的MS SQL Server版本为SQL Server 2005, SQL Server 2005是基于SQL Server 2000的强大功能之上推出的一个完整的数据管理和分析的解决方案。

SQL Server 数据平台包括以下工具：

关系型数据库:关系型数据库引擎，支持结构化和非结构化（XML）数据。

复制服务: 数据复制可用于数据分发、处理移动数据应用、系统高可用、企业报表解决方案的后备数据可伸缩存储、与异构系统的集成等，包括已有的Oracle数据库等。

通知服务: 用于开发、部署可伸缩应用程序的通知服务，能够向不同的连接和移动设备

发布及时的信息更新。

集成服务：可以支持数据仓库和企业范围内数据集成的抽取、转换和装载能力。

分析服务：联机分析处理（OLAP）功能可用于多维存储的大量、复杂的数据集的快速高级分析。

报表服务：可创建、管理和发布传统的、可打印的报表和交互的、基于Web的报表。

管理工具：SQL Server 包含的集成管理工具可用于高级数据库管理和调谐，构建于SQL Server内的内嵌Web service支持确保了和其他应用及平台的互操作能力。

开发工具：SQL Server 为数据库引擎、数据抽取、转换和装载（ETL）、数据挖掘、OLAP和报表提供了和Microsoft Visual Studio® 相集成的开发工具，以实现端到端的应用程序开发能力。SQL Server中每个主要的子系统都有自己的对象模型和API，能够以任何方式将数据系统扩展到不同的商业环境中。

安装SQL Server 2005的软硬件需求如下

最低要求	
处理器	Intel Pentium 或兼容的 166 MHz 或更高速度处理器。
操作系统	运行在 Microsoft Windows NT® Server 版本 4.0 Service Pack 5 (SP5) 或更高版本本、Microsoft Windows® 2000 Server、Microsoft Windows 2000 Advanced Server 和 Microsoft Windows 2003 Server
内存	64 MB RAM，建议使用 128 MB。
硬盘	95-270 MB 可用空间用于服务器，250 MB 用于典型安装。