Exadata 上的 Writeback 和 Writethrouth

第一篇 硬盘篇

(www.lunar2013.com)

以前写过一些关于 Exadata 上磁盘管理相关的文章：

Exadata 的磁盘自动管理-1-教你读懂 cell alert 中的磁盘相关信息
Exadata 的磁盘自动管理-2-Cell 上各种磁盘相关的概念
Exadata 的磁盘自动管理-3-磁盘自动管理的操作规则

如何看待 exadata 的 cell 节点出现的 writethrough/wirteback 模式更换或者控制器充放电信息

Exadata 更换硬盘的操作过程和解释

今天偶然间看见一段 alert 的信息，这是 Exadata 上 Disk Controller BBU 充放电的相关信息，具体解释请参见《如何看待 exadata 的 cell 节点出现的 writethrough/wirteback 模式更换或者控制器充放电信息》：

dm01cel01: 25_1 2014-01-17T02:00:52+08:00 info "The disk controller battery is executing a learn cycle and may temporarily enter WriteThrough Caching mode as part of the learn cycle. Disk write throughput might be temporarily lower during this time. The flash drives are not affected. The battery learn cycle is a normal maintenance activity that occurs quarterly and runs for approximately 1 to 12 hours. Note that many learn cycles do not require entering WriteThrough caching mode. When the disk controller cache returns to the normal WriteBack caching mode, an additional informational alert will be sent. Battery Serial Number : 13718 Battery Type : iBBU08 Battery Temperature : 42

C Full Charge Capacity : 1345 mAh Relative Charge : 100 % Ambient Temperature : 23 C"

dm01cel01: 25_2 2014-01-17T07:34:12+08:00 clear "All disk drives are in WriteBack caching mode. Battery Serial Number : 13718 Battery Type : iBBU08 Battery Temperature : 46 C Full Charge Capacity : 1341 mAh Relative Charge : 51 % Ambient Temperature : 23 C"
dm01cel01: 26 2014-01-20T10:49:03+08:00 info "This is a test trap"
dm01cel01: 27_1 2014-03-01T12:27:00+08:00 critical "Cell configuration check discovered the following problems: Check Exadata configuration via ipconf utility Verifying of Exadata configuration file /opt/oracle.cellos/cell.conf Checking DNS server on 0.48.0.10 : FAILED Error. Overall status of verification of Exadata configuration file: FAILED [INFO] The ipconf check may generate a failure for temporary inability to reach NTP or DNS server. You may ignore this alert, if the NTP or DNS servers are valid and available. [INFO] You may ignore this alert, if the NTP or DNS servers are valid and available. [INFO] As root user run /usr/local/bin/ipconf -verify -semantic to verify consistent network configurations."

dm01cel01: 27_2 2014-03-02T12:25:18+08:00 clear "The cell configuration check was successful."
dm01cel01: 28_1 2014-03-08T12:26:54+08:00 critical "Cell configuration check discovered the following problems: Check Exadata configuration via ipconf utility Verifying of Exadata configuration file /opt/oracle.cellos/cell.conf Checking DNS server on 10.48.0.10 : FAILED Error. Overall status of verification of Exadata configuration file: FAILED [INFO] The ipconf check may generate a failure for temporary inability to reach NTP or DNS server. You may ignore this alert, if the NTP or DNS servers are valid and available. [INFO] You may ignore this alert, if the NTP or DNS servers are valid and available. [INFO] As root user run /usr/local/bin/ipconf -verify -semantic to verify consistent network configurations."

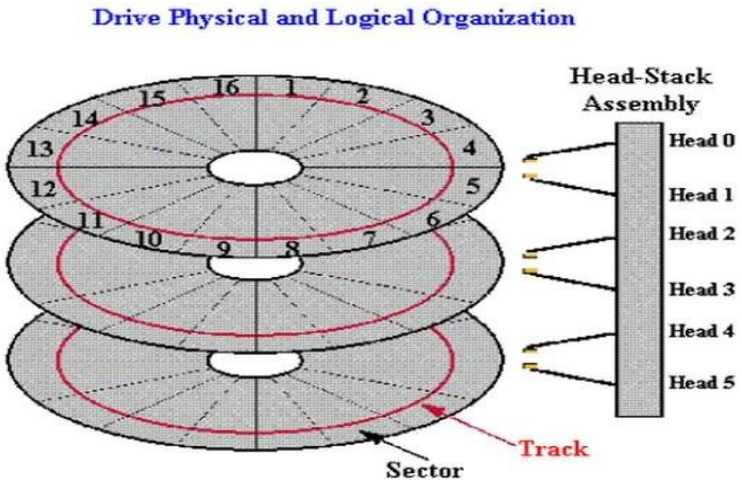
dm01cel01: 28_2 2014-03-09T12:25:19+08:00 clear "The cell configuration check was successful."

联想到 Exadata 上 Flash Card 的“Writeback 和 Writethrough”功能的变迁，以及目前为什么社会上各家一体机的架构（我所了解的大一点的 Oracle 运维三方基本都有了，没有的也在研发中了，只有一些目前宣称只做软件的三方还没做一体机，O(∩_∩)O 哈哈~），忽然间想更多的了解下各种硬盘和 SSD 的发展历程，于是就开始给自己扫盲……

从 Exadata 诞生的 V1，到现在的 Exadata X4，存储节点上使用过的硬盘有几种：

V1: 12 x 300 GB 15,000 RPM SAS or 12 x 1 TB 7,200 RPM SATA
V2: 12 x 600 GB 15,000 RPM SAS or 12 x 2 TB 7,200 RPM SATA
X2: 12 x 600 GB 15,000 RPM SAS or 12 x 2 TB 7,200 RPM SATA

X3: 12 x 600 GB 15,000 RPM High Performance disks or 12 x 3 TB 7,200 RPM High Capacity disks
X4: 12 x 1.2 TB 10,000 RPM High Performance disks or 12 x 4 TB 7,200 RPM High Capacity disks



从上边的标准 3.5 吋 HDD 拆解图片我们可以看到存储数据的盘片，一张这样的盘片容量可达几百 GB 甚至 1TB，比如目前希捷推出的单碟 1TB 系列酷鱼硬盘。可以说，磁盘存储技术的发展并不慢，不过仅限于存储密度方面，随着单位面积存储容量的提高，我们可以享用到更高容量的硬盘，但是读写数据的速度上并没有太大突破。

这是为什么呢？带着这个疑问，我开始了漫长的 Learn from Google and Learn from Google……

我本人是做软件的，对于硬件的东西其实不太了解，带着这个好奇，我在网上查找了很多的资料，给自己科普了一遍关于硬盘发展的一些知识，现在大概了解了一点点相关知识。脑子记忆力不好，每天一团浆糊，因此，好不容易花了很多时间研究，赶紧记录下来。当然，对于相当一部分做 IT 行业的朋友来说，甚至是其他一些玩游戏、玩电脑、攒机器的朋友来说，这些根本就不算是个菜，O(∩_∩)O 哈哈~。

硬盘接口是硬盘与主机系统间的连接部件，作用是在硬盘缓存和主机内存之间传输数据，不同的硬盘接口决定着硬盘与控制器之间的连接速度，因此了解一款磁盘阵列的硬盘接口往往是衡量这款产品的关键指标之一。

我第一次接触计算机是 90 年代初，高中的时候，用的是中华学习机，最近才知道那个小东西就是个类似小终端的东西，没有硬盘。那么回忆一下，我接触的第一台计算机，应该是大学期间的 286 了，那时应该也是第一次看到了 IDE 硬盘……

首先回顾一下相关大记事：

各种硬盘接口的相关大记事：
1，1984 年底：Compaq 开发出了 IDE 接口，至今，还有很多设备都在使用 IDE 接口。平常所说的 IDE 接口，也称之为 ATA 接口。

2，SCSI 最早是 1979 年由美国的 Shugart 公司（希捷公司前身）制订的，在 1986 年获得了 ANSI（美国标准协会）的承认,目前 SCSI 接口的最大应用 SCSI 硬盘已经被 SAS 接口硬盘取代，这也是这 2-3 年内发生的事

3，SATA 接口的历史： 2001 年，由 Intel、Dell、IBM、希捷、迈拓这几大厂商组成的 Serial ATA 委员会正式确立了 Serial ATA 1.0 规范。2002 年，虽然串行 ATA 的相关设备还未正式上市，但 Serial ATA 委员会已抢先确立了 Serial ATA 2.0 规范。 而 SATA 产品真正的普及也就是这 3-4 年的事！

各种盘发展的相关大记事：

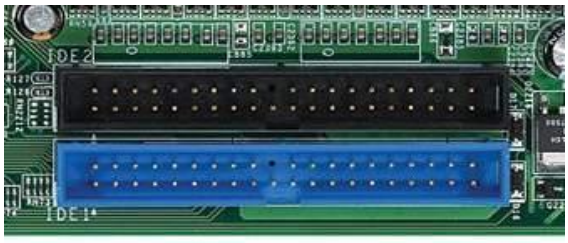
1956 年，IBM 公司发明了世界上第一块硬盘。
1968 年，IBM 重新提出“温彻斯特”（Winchester）技术的可行性，奠定了硬盘发展方向。
1970 年，StorageTek 公司(Sun StorageTek)开发了第一个固态硬盘驱动器。
1989 年，世界上第一款固态硬盘出现。
2006 年 3 月，三星率先发布一款 32GB 容量的固态硬盘笔记本电脑，
2007 年 1 月，SanDisk 公司发布了 1.8 寸 32GB 固态硬盘产品，3 月又发布了 2.5 寸 32GB 型号。
2007 年 6 月，东芝推出了其第一款 120GB 固态硬盘笔记本电脑。
2008 年 9 月，忆正 MemoRight SSD 的正式发布，标志着着中国企业加速进军固态硬盘行业。
2009 年，SSD 井喷式发展，各大厂商蜂拥而来，存储虚拟化正式走入新阶段。
2010 年 2 月，镁光发布了全球首款 SATA 6Gbps 接口固态硬盘，突破了 SATAII 接口 300MB/s 的读写速度。
2012 年，苹果公司在笔记本电脑上应用容量为 512G 的固态硬盘。[1]
2012 年 7 月 ，Goldendisk 深圳云存科技推出全球第一款体积最小的 CFast 固态硬盘。

IDE 盘

IDE（Integrated Drive Electronics），即“电子集成驱动器”，也是一种极为常用的接口，它的本意是指把“硬盘控制器”与“盘体”集成在一起的硬盘驱动器 ，人们也习惯用 IDE 来称呼最早出现 IDE 硬盘。。
IDE 的工作方式需要 CPU 的全程参与，CPU 读写数据的时候不能再进行其他操作。

ATA（Advanced Technology Attachment ）接口标准是 IDE(Integrated Drive Electronics)硬盘的特定接口标准，目前有：ATA-1 、 ATA-2 、 ATA-3 、 ATA-4 、 ATA-5 、 ATA-6 、 ATA-7
这类 ATA 接口协议都是并行 ATA(Paralle ATA)接口协议。PATA 接口一般使用 16-bit 数据总线， 每次总线处理时传送 2 个字节。

ATA-7 是 ATA 接口的最后一个版本，也叫 ATA133 。只有迈拓公司推出一系列采用 ATA133 标准的硬盘，这是第一种在接口速度上超过 100MB/s 的 IDE 硬盘。
迈拓是目前惟一一家推出这种接口标准硬盘的制造商，而其他 IDE 硬盘厂商则停止了对 IDE 接口的开发，转而生产 Serial ATA 接口标准的硬盘。
ATA133 接口支持 133 MB/s 数据传输速度，在 ATA 接口发展到 ATA100 的时候，这种并行接口的电缆属性、连接器和信号协议都表现出了很大的技术瓶颈，而在技术上突破这些瓶颈存在相当大的难度。



IDE 接口

IDE 硬盘

IDE 硬盘都有跳线

SATA 盘（串行 ATA）

随着 CPU 时钟频率和内存带宽的不断提升，PATA 逐渐显现出不足来。串行总线接口协议(Serial ATA，SATA)应运而生。

SATA 的全称是 Serial Advanced Technology Attachment，是 由 Intel、IBM、Dell、APT、Maxtor 和 Seagate 公司共同提出的硬盘接口规范 。

SATA 接口需要硬件芯片的支持， ，例如 Intel ICH5(R)、VIA VT8237、nVIDIA 的 MCP RAID 和 SiS964，如果主板南桥芯片不能直接支持的话，就需要选择第三方的芯片 ，例如 Silicon Image 3112A 芯片等，不过这样也就会产生一些硬件性能的差异，并且驱动程序也比较繁杂

从外型上，SATA 盘和 IDE 盘没什么区别。但是 SATA 以它串行的数据发送方式得名。在数据传输的过程中，**数据线和信号线独立使用，并且传输的时钟频率保持独立**，因此同以往的 PATA 相比，SATA 的传输速率可以达到并行的 30 倍。
可以说:SATA 技术并不是简单意义上的 PATA 技术的改进，而是一种全新的总线架构。目前有 SATA-1 和 SATA-2 两种标准。

尽管 SATA 在诸多性能上远远优越于 PATA，甚至在某些单线程任务的测试中，表现出了不输于 SCSI 的性能，然而它的机械底盘仍然为低端应用设计的，在面对大数据吞吐量或者多线程的传输任务时，相比 SCSI 硬盘，仍然显得力不从心。

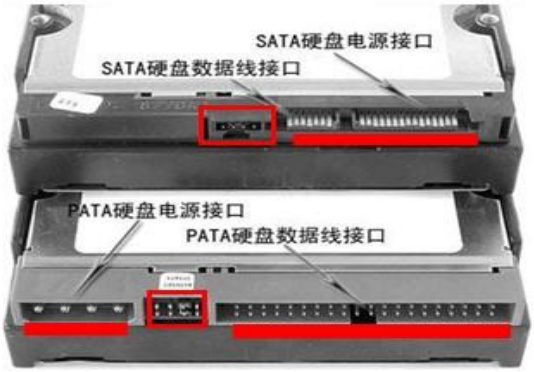
SATA 接口需要硬件芯片的支持 ，例如 Intel ICH5(R)、VIA VT8237、nVIDIA 的 MCP RAID 和 SiS964，如果主板南桥芯片不能直接支持的话，就需要选择第三方的芯片。

我们有时候把 IDE 硬盘和 SATA 硬盘统称为 ATA 硬盘，即 PATA（Parallel ATA，即 IDE 盘）和 SATA（Serial ATA，即 SATA 盘）。他们都采用 IDE 插槽与系统连接。

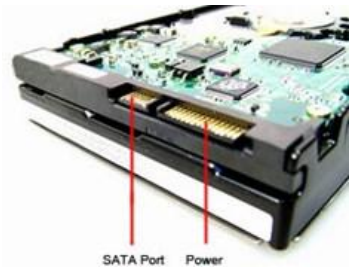
主板 SATA 接口

直连 SATA 硬盘

eSATA 接口移动硬盘



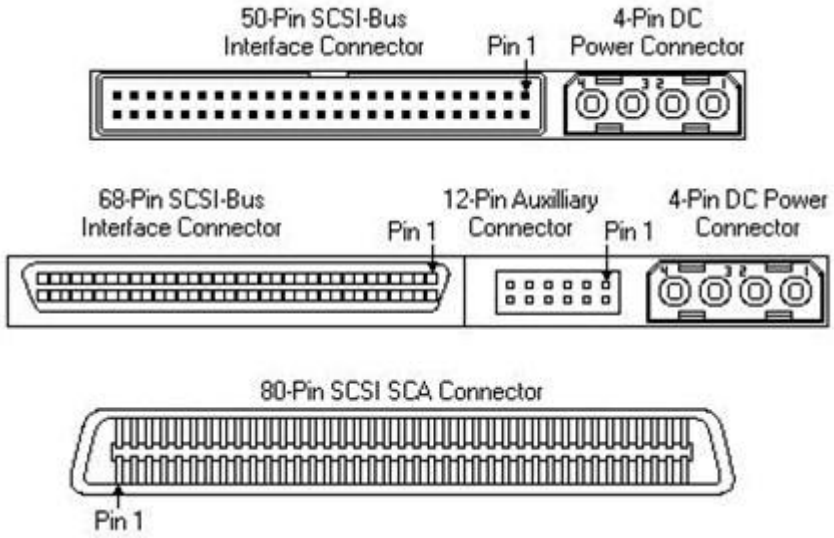
SATA 盘无需跳线



了解 SCSI 盘之前，先了解下 SCSI 接口

SCSI，小型计算机系统接口（英语：Small Computer System Interface; 简写：SCSI），一种用于计算机和智能设备之间（硬盘、软驱、光驱、打印机、扫描仪等）系统级接口的独立处理器标准。 SCSI 是一种智能的通用接口标准。
它是各种计算机与外部设备之间的接口标准。SCSI 是个多任务接口，设有母线仲裁功能。挂在一个 SCSI 母线上的多个外设可以同时工作。
SCSI 上的设备平等占有总线。SCSI 接口可以同步或异步传输数据。

其实 SCSI 并不是专为硬盘设计的，实际上它是一种总线型接口。独立于系统总线工作。



各种 SCSI 接口：

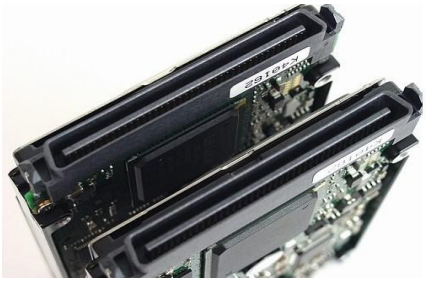
SCSI 盘

一般而言，ATA 硬盘采用 IDE 插槽与系统连接，而每 IDE 插槽即占用一个 IRQ(中断号)，而每两个 IDE 设备就要占用一个 IDE 能道，虽然附加 IDE 控制卡等方式可以增加所支持的 IDE 设备数量，但总共可连接的 IDE 设备数最多不能超过 15 个。
而 SCSI 的所有设备只占用一个中断号(IRQ)，因此它支持的磁盘扩容量要比 ATA 更为巨大。

发送命令到一个 SCSI 设备，磁盘可以移动驱动臂定位磁头，在磁盘介质和缓存中传递数据，整个过程在后台执行。
也就是说，对于 SCSI 而言，它有独立的芯片负责数据处理，当 CPU 将指令传输给 SCSI 后，随即去处理后续指令，其它的相关工作就交给 SCSI 控制芯片来处理，当 SCSI“处理器”处理完毕后，再次发送控制信息给 CPU，CPU 再接着进行后续工作。
因此，SCSI 硬盘 CPU 占用率低、并行处理能力强。这样可以同时发送多个命令同时操作，适合大负载的 I/O 应用。在磁盘阵列上的整体性能也大大高于基于 ATA 硬盘的阵列。

SCSI 在传输速率和容错性上有极好的表现，但是它昂贵的价格使得用户望而却步。而下一代 SCSI 技术 SAS 的诞生，则更好的兼容了性能和价格双重优势。

SCSI 硬盘



SCSI 硬盘

SAS 盘（串行 SCSI）——主流硬盘

SAS 是 Serial Attached SCSI 的缩写，即串行连接 SCSI。2001 年 11 月 26 日，Compaq、IBM、LSI 逻辑、Maxtor 和 Seagate 联合宣布成立 SAS 工作组，其目标是定义一个新的串行点对点的企业级存储设备接口。

SAS 相对于 SCSI 技术而言，也是一种革命性的变革。它既利用了已经在实践中验证的 SCSI 功能与特性，又以此为基础引入了 SAS 扩展器。使 SAS 系统可以连接更多的设备，其中每个扩展器允许连接多个端口，每个端口可以连接 SAS 设备、主机或其他 SAS 扩 展器。

对于 SCSI 系统而言，它有独立的芯片负责数据处理，当 CPU 将指令传输给 SCSI 后，随即去处理后续指令，其它的相关工作就交给 SCSI 控制芯片来处理。当 SCSI“处理器”处理完毕后，再次发送控制信息给 CPU，CPU 再接着进行后续工作，因此 SCSI 系统对 CPU 的占用率很低。

因此，相对于 SATA 盘使用 CPU 来处理磁盘系统和计算系统之间数据流而言，SCSI 硬盘使用 SCCI 控制器来完成这一工作，这样就允许一个用户对其进行数据传输的同时，另一位用户同时对其进行数据查找，这就是 SCSI 硬盘并行处理能力的体现。

SAS 的接口技术可以向下兼容 SATA。SAS 系统的背板(Backplane)既可以连接具有双端口、高性能的 SAS 驱动器，也可以连接高容量、低成本的 SATA 驱动器。因此很多存储上 SAS 盘和 SATA 盘可以共存。但是，SATA 系统并不兼容 SAS，所以 SAS 驱动器不能连接到 SATA 背板上。

这主要是因为，在物理层，SAS 接口和 SATA 接口完全兼容，SATA 硬盘可以直接使用在 SAS 的环境中，从接口标准上而言，SATA 是 SAS 的一个子标准，因此 SAS 控制器可以直接操控 SATA 硬盘，但是 SAS 却不能直接使用在 SATA 的环境中，因为 SATA 控制器并不能对 SAS 硬盘进行控制；在协议层，SAS 由 3 种类型协议组成，根据连接的不同设备使用相应的协议进行数据传输。其中串行 SCSI 协议(SSP)用于传输 SCSI 命令；SCSI 管理协议(SMP)用于对连接设备的维护和管理；SATA 通道协议(STP)用于 SAS 和 SATA 之间数据的传输。因此在这 3 种协议的配合下，SAS 可以和 SATA 以及部分 SCSI 设备无缝结合。

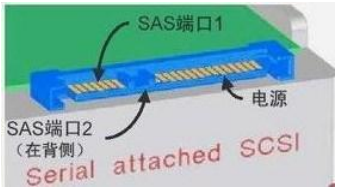
由于 SAS 由 SCSI 发展而来，在主机端会有众多的厂商兼容。SAS 采用了点到点的连接方式，每个 SAS 端口提供 3Gb 带宽，传输能力与 4Gb 光纤相差无几，这种传输方式不仅提高了高可靠性和容错能力，同时也增加了系统的整体性能。

在磁盘端，SAS 协议的交换域能够提供 16384 个节点，而光纤环路最多提供 126 个节点。

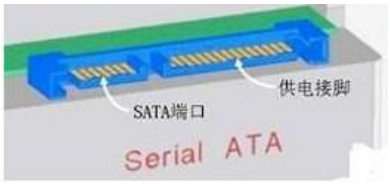
如果把 SCSI 和 SAS 进行对比，除了速度上的不同之外，相比与 SCSI，SAS 有一个非常突出的优势。在 SCSI 技术中，不同类型的设备是连接成一个链，所有的设备都按照最慢的一个

设备的速度运行。而在 SAS 技术中，情况不再是这样。即使是不同类型的设备，每个设备都可以按照自己的速度运行。

SAS 产品的成本从芯片级开始，都远远低于 FC，而正是因为 SAS 突出的性价比优势，使 SAS 在磁盘接口领域，给光纤存储带来极大的威胁。



SAS 接口



SATA 接口

SAS 背板插座

FC（Fibre Channel ，光纤盘）

FC 硬盘是指采用 FC-AL(Fiber Channel Arbitrated Loop，光纤通道仲裁环）接口模式的磁盘，因此起名为光纤硬盘，现在也支持铜线物理通道。
就像是 IEEE-1394, Fibre Channel 实际上定义为 SCSI-3 标准一类，属于 SCSI 的同胞兄弟。

通过光学连接设备最大传输距离可以达到 10KM。通过 FC-loop 可以连接 127 个设备，也就是为什么基于 FC 硬盘的存储设备通常可以连接几百颗甚至千颗硬盘提供大容量存储空间。
FC-AL 使光纤通道能够直接作为硬盘连接接口，为高吞吐量性能密集型系统的设计者开辟了一条提高 I/O 性能水平的途径。目前高端存储产品使用的都是 FC 接口的硬盘。

FC 盘的缺点在于，FC 磁盘通道工作于环路模式下，一个光纤环路在同一时间只能实现单个磁盘的 I\O，导致 FC 带宽不能被充分利用，并且影响到磁盘并行访问的性能。

解决方案	应用	附注
Fiber Channel	在线，高可用性，随机读取	适用于型企业中的关键任务资料的存储，例如SAN，最多支持1600万个位址，线缆最长可达10公里，相当昂贵。
Serial Attached SCSI（SAS）	在线，高可用性，随机读取	适用于大、中型企业关键任务资料的存储，效能高而且在本端层次上的扩充性极高，比FC便宜，与SATA兼容

	4Gb FC	SAS
协议带宽	全双工 4Gb/s	全双工 3Gb/s
扩展能力	每环路126设备	每交换域16，384设备
连接机制	共享环路带宽	交换式结构
连接距离	10公里	8米
主机连接	每线缆400MB/s	每线缆1200MB/s

总结：

1，IDE 硬盘和 SATA 硬盘统称为 ATA 硬盘，即 PATA（Parallel ATA，即 IDE 盘）和 SATA（Serial ATA，即 SATA 盘）。他们都采用 IDE 插槽与系统连接。

SATA 硬盘主要是应用于 PC 机上面的，对于低端的小型服务器应用来说，可以采用最新的 SATA 硬盘和控制器。

SATA 盘和 IDE 盘相比，前者数据线和信号线独立使用，并且传输的时钟频率保持独立，因此 SATA 的传输速率能达到 PATA 的 30 倍。

2，SCSI 盘（Parallel SCSI）和 SAS 盘（Serial Attached SCSI）都采用 SCSI 接口连接。带宽，稳定性和传输速率都远远超过 SATA 盘，且 CPU 占用率低，并行处理能力强。

与 SATA 盘相比，SCSI/SAS 具有 CPU 占用率低（由 SCSI 控制器完成一部分原本 CPU 的工作），多任务并发操作效率高，连接设备多，连接距离长等优点。SCSI/SAS 硬盘的价格较贵，同样容量的 SCSI/SAS 硬盘价格会比 SATA 硬盘贵 80%以上

3，在 SCSI 技术中，不同类型的设备是连接成一个链，所有的设备都按照最慢的一个设备的速度运行。

而在 SAS 技术中，情况不再是这样。即使是不同类型的设备，每个设备都可以按照自己的速度运行。

SATA FC SAS三种硬盘的比较

解决方案	应用	附注
Fibre Channel	在线，高可用性，随机读取	适用于型企业中的关键任务资料的存储，例如SAN，最多支持1600万个位址，线缆最长可达10公里，相当昂贵。
Serial Attached SCSI(SAS)	在线，高可用性，随机读取	适用于大、中型企业关键任务资料的存储，效能高而且在本端层次上的扩充性极高，比FC便宜，与SATA兼容
SATA	在线、近线作业，高可用性，随机读取，循序读取	容里高、成本低，与SAS兼容
Tape		容里高，适合于保存和保证offsite资料安全

对于用户来说：单纯比较硬盘并不一定是越贵的越好，关键是看是否适合自己的应用。另外单纯硬盘的硬盘接口协议也不是衡量一个存储系统性能指标的唯一要素，除了硬盘性能指标以外，存储系统的硬件设计，前端主机接口等性能指标也同样对存储系统的整体性能影响巨大。