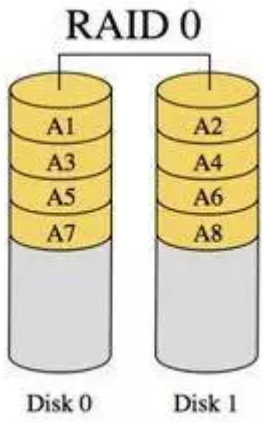
第4章 RAID技术详解

RAID 类型介绍

原创 EMC中文技术社区 戴尔易安信技术支持 *2016-05-03*

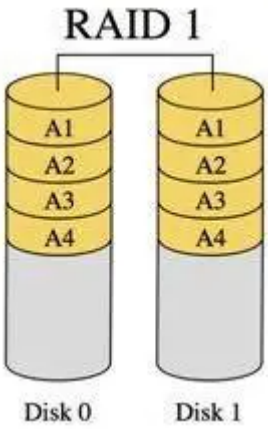
RAID (Redundant Array of Independent/InexpensiveDisks)，独立磁盘冗余阵列，是一种将多块独立的硬盘（物理硬盘）按不同的组合方式形成一个硬盘组（逻辑硬盘），从而提供比单块硬盘更大的存储容量、更高的可靠性和 更快的读写性能等。该概念最早由加州大学伯克利分校的几名教授于1987年提出。早期主要通过RAID控制器等硬件来实现RAID磁盘阵列，后来出现了基 于软件实现的RAID，比如mdadm等。按照磁盘阵列的不同组合方式，可以将RAID分为不同级别，包括RAID0到RAID6等7个基本级别，以及 RAID0+1和RAID10等扩展级别。不同RAID级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



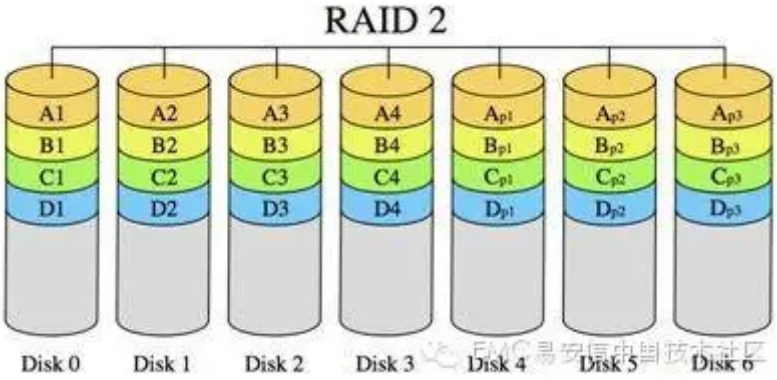
**
**

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。



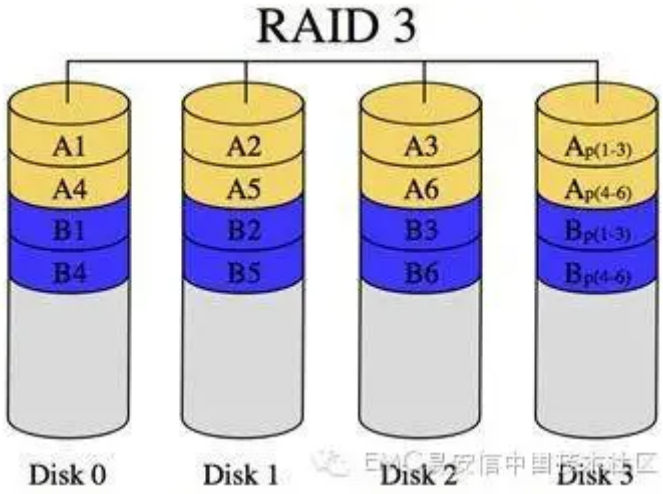
**
**

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。

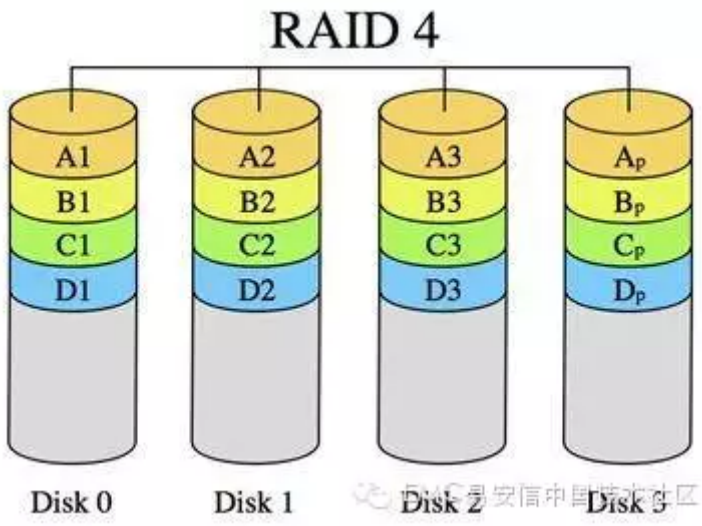


**
**

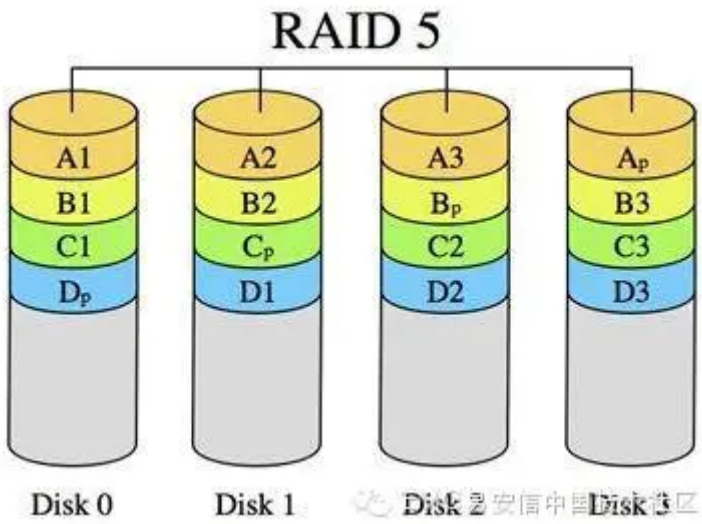
RAID 3：与RAID 2类似，不同的是：1）以字节（byte）为单位进行—条带化Ⅱ处理；2）以奇偶校验码取代海明码。RAID3的读写性能都还不错，而且存储利用率也相当高，可达到(n-1)/n。但是对于随即读写操作，奇偶盘会成为写操作的瓶颈。



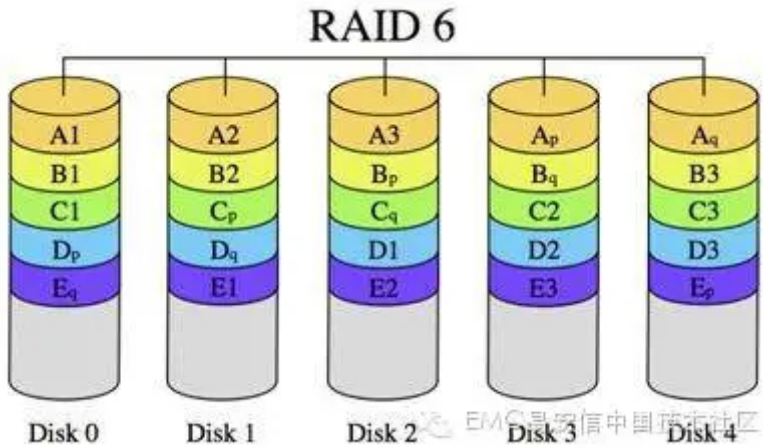
RAID 4：与RAID 3的分布结构类似，不同的是RAID 4以数据块（block）为单位进行奇偶校验码的计算。另外，与RAID2和RAID3不同的是，RAID4中各个磁盘是独立操作的，并不要求各个磁盘的磁头同步转动。因此，RAID4允许多个I/O请求并行处理。



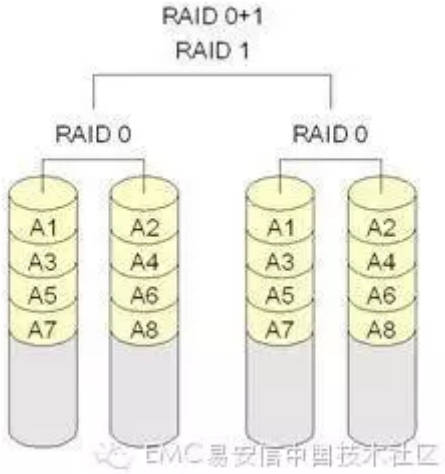
RAID 5: RAID 3和RAID 4都存在同一个问题，就是奇偶校验码放在同一个硬盘上，容易造成写操作的瓶颈。RAID5与RAID4基本相同，但是其将奇偶校验码分开存放到不同的硬盘上去，从而减少了写奇偶校验码带来瓶颈的可能性。



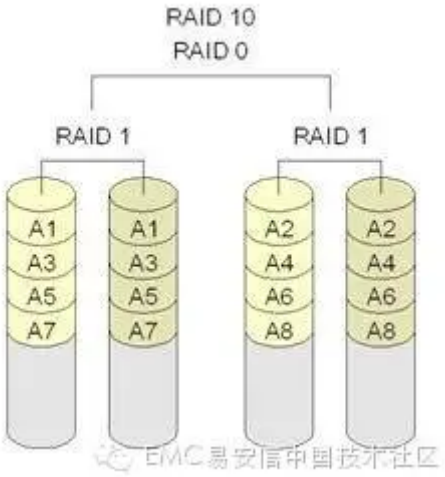
RAID 6：在RAID 5的基础上，RAID 6又另外增加了一组奇偶校验码，从而获得更高的容错性，最多允许同时有两块硬盘出现故障。但是，新增加的奇偶校验计算同时也带来了写操作性能上的损耗。



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的方式组成一条带化的硬盘阵列，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。



存储基础知识 - 磁盘寻址

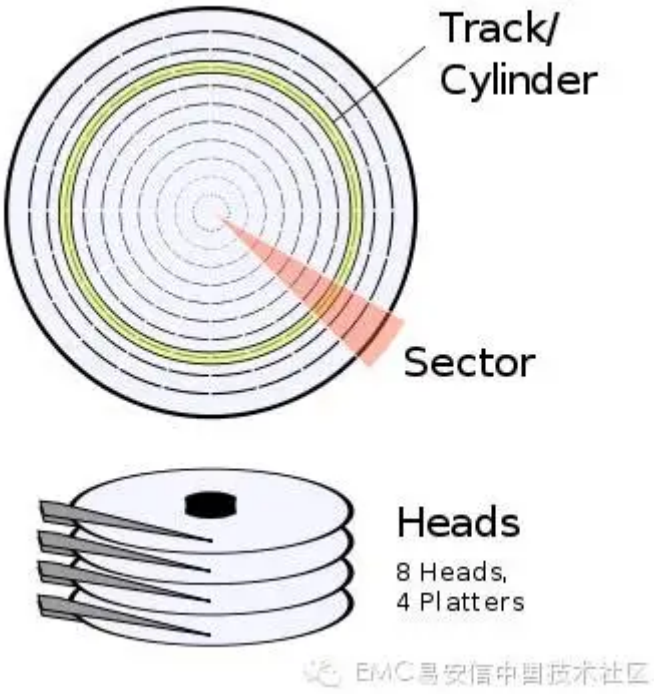
原创 EMC中文技术社区 戴尔易安信技术支持 *2016-04-30*

1. 磁盘驱动器

维基百科网址：http://en.wikipedia.org/wiki/Hard_disk_drive

磁盘物理结构图如下：

磁盘逻辑组成图如下：



2. 什么是CHS（cylinder head sector）

维基百科网址：http://en.wikipedia.org/wiki/Cylinder-head-sector

通过上面材料，我们了解到磁盘通常由多个盘片、多个磁头组成。

每个盘片对应一个磁头（head），每个盘片被化成多个同心圆(track/cylinder)，每个同心圆被切断成多个段（sector）。磁盘存储最小单位是sector，那么如何对sector进行定位？

CHS是早期在IBM PC架构上面用来进行磁盘寻址的办法。

CHS是一个三元组，组成如下：

- 一共24个 bit位。
- 其中前10位表示cylinder，中间8位表示head，后面6位表示sector。
- 最大寻址空间

随着科技大发展，磁盘容量大幅提升。远远超过了8GB寻址范围，如何对8GB之外空间进行寻址？历史上曾经CHS从24位扩展到多28位，实现寻址128GB，但是面对现在磁盘2TB容量还是无能为力，下面我们请出最终解决方案LBA。

3. 什么是LBA（logical block addressing)

维基百科网址：http://en.wikipedia.org/wiki/Logical_Block_Addressing#CHS_conversion

正如上文所说，LBA是用来取代CHS。那么LBA是怎么实现磁盘寻址？

- LBA是一个整数，通过转换成CHS格式完成磁盘具体寻址。
- LBA采用48个bit位寻址，最大寻址空间128PB。

LBA与CHS转换规则是怎么样的？

CHS->LBA

$$LBA = ((C \times HPC) + H) \times SPT + S - 1$$

LBA->CHS

$$C = LBA \div (SPT \times HPC)$$
$$H = (LBA \div SPT) \bmod HPC$$
$$S = (LBA \bmod SPT) + 1$$

小结：

不管CHS也好，还是LBA也好。磁盘存储寻址都需要通过cylinder、head、sector来实现；CHS、LBA都是一个数字，CHS按照固定格式把24个bit位分成成cylinder、head、sector；LBA则需要通过求模运算得出cylinder、head、sector。

如何计算磁盘性能

原创 EMC中文技术社区 戴尔易安信技术支持 *2016-05-10*

磁盘是主要的物理存储设备，机械硬盘的性能受其转速（RPM），寻道时间（Seek Time）以及旋转延迟（Rotational latency）的影响，本文将描述如何通过这些参数计算磁盘的性能。

磁盘规格

机械硬盘的性能指标有三个重要的参数：

- 寻道时间 – 在磁道之间移动磁头所花费的时间
- 旋转延迟 – 盘片将数据旋转至磁头下的时间
- 传输速率 – 磁盘的带宽

理解这些参数之间的关系有助于了解一块磁盘的性能，这些值在决定磁盘性能的两个基本度量的时候非常有用：吞吐量和响应时间

寻道时间

寻道时间以毫秒（ms）来计算，不同磁盘的寻道时间不同。平均寻道时间是经常使用的度量，对于一块15k rpm的3.5英寸SAS盘，其平均寻道时间是3.8ms。减少磁盘寻道所花费的时间能增强性能。i/o类型也会影响寻道时间，连续i/o拥有最少的寻道时间，因为读写头可以在盘片上连续操作，而随机i/o就相对有较长的寻道时间，因为磁头始终需要在不同的磁道间切换。

延迟

延迟以毫秒（ms）来计算，更高转速的磁盘其延迟更小。下表显示了不同转速的磁盘所对应的延迟：

传输速率

传输速率以MB/s来计算，它又可以进一步分为内部/外部速率。内部速率是指在盘片上读写数据的快慢，盘片外圈速率要高于盘片里圈，而且对于同样的线性距离，也拥有更多的扇区。比如对于一个使用连续带宽的应用，3.5-inch 15k rpm SAS磁盘可以提供50MB/s的内圈速率以及100MB/s的外圈速率。

外部传输速率是指磁盘的连线头到HBA或NIC的传输速率。厂商通常给出的都是突发速率，且假定是内部连接（DAS）。对于存储系统来说，比如VNX，同一个RAID组内的磁盘是共享后端此部分速率的，因此通常达不到厂商给出的突发速率。存储系统的总线架构，实际传输速率更多是由后端传输协议、仲裁时间以及后端端口容量来决定的。

计算平均响应时间

平均响应时间是指一个请求从排队开始一直到执行结束所花费的时间，计算公式为：响应时间 = （队列长度+1）*平均响应时间

比如，某块磁盘的平均响应时间为6ms，队列长度为6，那么响应时间 = 42ms = (6+1)*6 ms

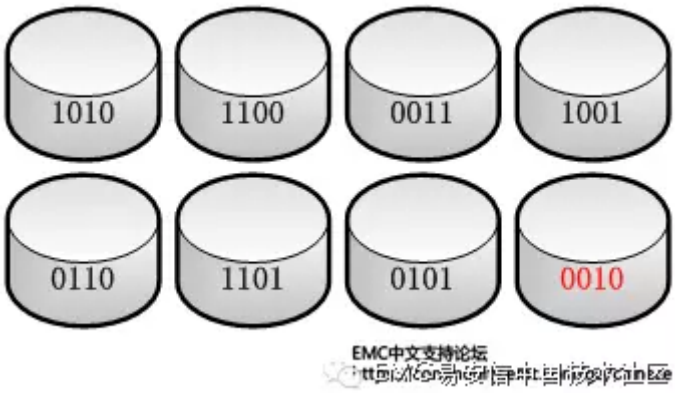
浅谈RAID写惩罚（Write Penalty）与IOPS计算

原创 EMC中文技术社区 戴尔易安信技术支持 *2016-05-12*

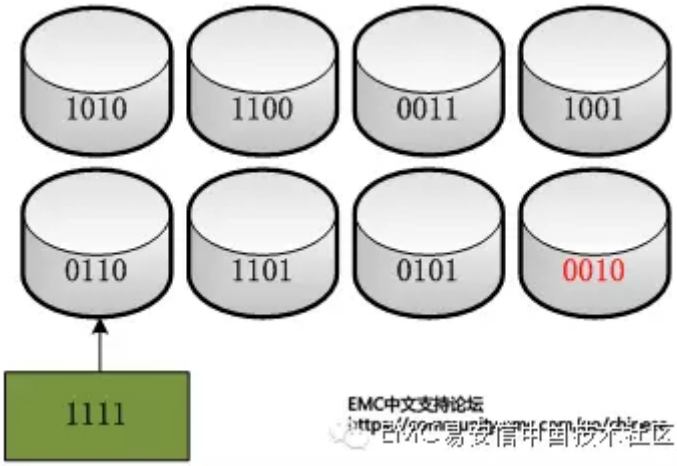
通常在讨论不同RAID保护类型的性能的时候，结论都会是RAID-1提供比较好的读写性能，RAID-5读性能不错，但是写入性能就不如RAID-1，RAID-6保护级别更高，但写性能相对更加差，RAID10是提供最好的性能和数据保护，不过成本最高等等。其实决定这些性能考虑的因素很简单，它就是RAID Write Penalty（写惩罚）。本文从原理上解释了不同RAID保护级别的写惩罚，以及通过写惩罚计算可用IOPS的方法。

RAID-5 Write Penalty的例子:

存储方案规划的过程中，最基本的考虑因素有两个，性能和容量。性能上的计算看可以分为IOPS和带宽需求。计算IOPS，抛开存储阵列的缓存和前端口不谈。计算后端物理磁盘的IOPS不能简单的把物理磁盘的最大IOPS相加而获得。原因是，对于不同的RAID级别，为了保证当有物理磁盘损坏的情况下可以恢复数据，数据写入的过程中都需要有一些特别的计算。比如对于RAID-5，条带上的任意磁盘上的数据改变，都会重新计算校验位。如下图所示，一个7+1的RAID-5的条带中，七个磁盘存储数据，最后一个磁盘存储校验位。



对于一个数据的写入，我们假设在第五个磁盘上写入的数据为1111，如下图所示。那么整个RAID-5需要完成写入的过程分为以下几步：



1. 读取原数据0110，然后与新的数据1111做XOR操作： 0110 XOR 1111 = 1001
2. 读取原有的校验位0010
3. 用第一步算出的数值与原校验位再做一次XOR操作： 0010 XOR 1001 = 1011
4. 然后将1111新数据写入到数据磁盘，将第三步计算出来的新的校验位写入校验盘。

由上述几个步骤可见，对于任何一次写入，在存储端，需要分别进行两次读+两次写，所以说RAID-5的Write Penalty的值是4。

不同RAID级别的Write Penalty：

下表列出了各种RAID级别的Write Penalty值：

RAID	Write Penalty
0	1
1	2
5	4
6	6
10	2

RAID-0：直接的条带，数据每次写入对应物理磁盘上的一次写入

RAID-1和10：RAID-1 和RAID-10的写惩罚很简单理解，因为数据的镜像存在的，所以一次写入会有两次。

RAID-5：RAID-5由于要计算校验位的机制存在，需要读数据、读校验位、写数据、写校验位四个步骤，所以RAID-5的写惩罚值是4。

RAID-6：RAID-6由于有两个校验位的存在，与RAID-5相比，需要读取两次校验位和写入两次校验位，所以RAID-6的写惩罚值是6。

计算IOPS:

根据上文的描述，在实际存储方案设计的过程中，计算实际可用IOPS的过程中必须纳入RAID的写惩罚计算。计算的公式如下：

物理磁盘总的IOPS = 物理磁盘的IOPS × 磁盘数目

可用的IOPS = （物理磁盘总的IOPS × 写百分比 ÷ RAID写惩罚） + （物理磁盘总的IOPS × 读百分比）

假设组成RAID-5的物理磁盘总共可以提供500 IOPS，使用该存储的应用程序读写比例是50%/50%，那么对于前端主机而言，实际可用的IOPS是：

（500 ×50% ÷ 4） + (500 * 50%) = 312.5 IOPS

具体不同应用程序的读写比例，可以参考：关于不同应用程序存储IO类型的描述

企业级闪存盘的结构和特征

原创 EMC中文技术社区 戴尔易安信技术支持 *2016-05-11*

随着信息的增长，存储用户的业务应用对于性能的需求越来越高。传统的使用增加磁盘数量的方式已经无法满足用户的I/O需求，因此企业级闪存盘得到了更加广泛的应用。本文将介绍企业级闪存盘的原理。

闪存盘也被称为固态硬盘，是新一代的磁盘。闪存盘采用了和传统机械磁盘相同的外形和连接器，以实现在存储机柜中使用闪存盘替代传统的机械磁盘。闪存盘使用基于半导体的固态存储来存取数据。与传统的机械磁盘相比，闪存盘不含移动部件，因此也没有寻道时间和旋转延迟。另外，因为是基于半导体的设备，闪存盘比机械磁盘更省电。闪存盘的关键组件包括控制器、I/O接口、大容量存储和缓存。控制器用于控制闪存盘的运行。I/O接口为闪存盘提供电力和数据访问。大容量存储由一组用于存储数据的非易失性NAND闪存芯片组成。缓存为数据提供临时存储空间。

闪存盘拥有极高的性能，能满足性能敏感型应用的需求。它拥有多个用于数据访问的并行I/O通道，一般来说，闪存芯片和通道的数量越多，闪存盘的内部带宽越大，其性能越高。闪存中的存储芯片是以块和页为逻辑单位组织的。页是闪存盘上能被读写的最小对象。页组合成块。每个块可以有32、64或者128个页。页的大小没有唯一标准，常见的大小有4KB，8KB和16KB。闪存盘的读操作发生在页层级，而写操作则发生在块层级。闪存盘模拟物理磁盘的逻辑块地址，每页占据一系列连续的物理块。举例来说，一个大小为4KB的页占据8个连续的512字节数据块。闪存盘特备适合那些文件块较小，随机读取工作较多，且要求响应时间持续保持较低水平的应用。那些需要快速处理大量数据或者实时数据处理的应用在使用闪存盘后能大幅提高性能。

一般，企业级闪存盘的吞吐量是传统机械磁盘的几十倍，而响应时间不到机械磁盘的十分之一。此外，每存储1TB数据，使用闪存盘比使用机械磁盘相比，最多可节省30%以上的电能。换算成单个I/O消耗的电能，闪存盘比机械磁盘节省90%以上。企业级闪存盘拥有以下主要特征：

NAND闪存技术：NAND设备采用不良块追踪技术和校验码，以保证数据一致性，实现最快的写入速度。

基于单级单元：NADN技术有二种单元设计方式。多层单元可以记录多个状态，所有每个单元可以存储多位数据。而单层单元的每个单元只存储一位数据。单层单元性能好和寿命长，适合企业级数据应用。

写入平衡技术：该技术保证经常更新的数据写入不同的位置，避免对同一单元的使用过于频繁。

Raid-7小七的故事

原创 EMC中文技术社区 戴尔易安信技术支持 *2016-05-14*

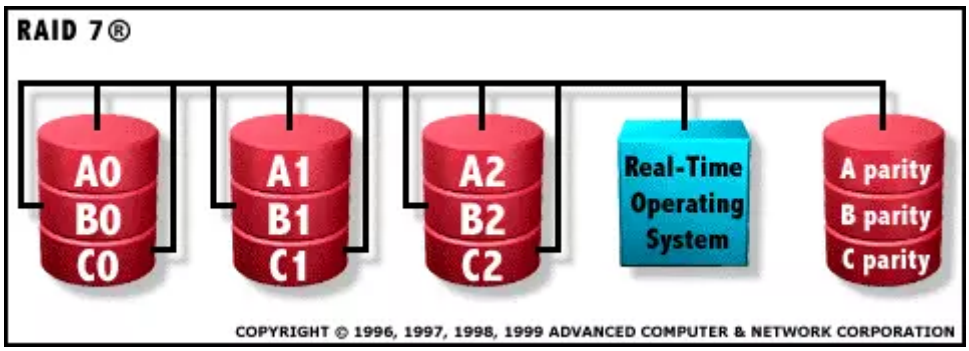
RAID是常见的一种数据保护方式，主要作用原理有两种：数据镜像和奇偶校验。最常见的raid类型有0，1，5，6。

今天为大家介绍一种比较罕见的类型 - RAID-7。

与其他RAID类型不同：

1. RAID-7是一个公司（Storage Computer Corporatio）的专利。
2. RAID-7是一个完整的存储阵列。RAID-7有自身的实时操作系统用来管理阵列。

RAID-7架构示意图如下：



1. 物理上RAID-7主要包括两部分：一个运行实时操作系统的控制器；二.多个Channel磁盘组。
2. 逻辑上 RAID-7类似于RAID3和RAID4。磁盘分布于多个Channel，一个Channel包含一组磁盘，校验盘可以分布于任意Channel。Channel之间通过X-BUS连接。
3. 异步IO。IO读写操作以及奇偶校验都直接在缓存里面完成。控制器负责数据从缓存写入磁盘。
4. 可以根据需求，将部分磁盘配置为Hot Standy模式。

5. 提供SNMP远程监控管理功能。

优点：

RAID-7所有IO操作都是异步发生在缓存里面，读写性能非常好，IO延迟低。模块化设计，扩展性强。

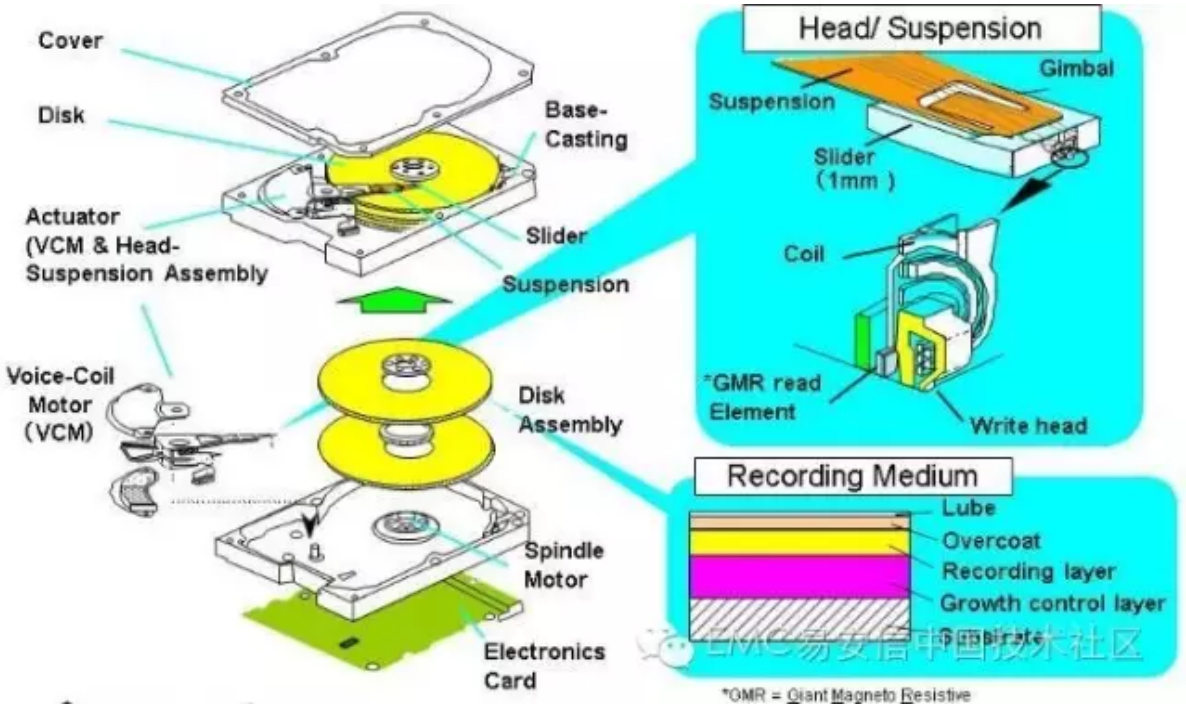
缺点：

单一厂商解决方案。成本高。数据保存在缓存里面，必须要求UPS防止数据丢失。

浅谈硬盘构造及IOPS的计算

原创 EMC中文技术社区 戴尔易安信技术支持 *2016-05-15*

在EMC存储设备中最基本的单元是硬盘，对于性能有重要影响的也是硬盘，硬盘本质是一种机械装置，在这里我们简单谈谈硬盘的构造。硬盘是由盘体、磁头、电机、硬盘控制器组成。下面我们分别介绍硬盘的各种单元：



盘体一般由多个盘片组成，这些盘片重叠在一起放在一个密封的盒中。硬盘的盘片是将磁粉附着在圆盘片的表面上.这些磁粉被划分成称为磁道的若干个同心圆，在每个同心圆的磁道上就好像有无数的任意排列的小磁铁，它们分别代表着0和1的状态。当这些小磁铁受到来自磁头的磁力影响时，其排列的方向会随之改变。利用磁头的磁力控制指定的一些小磁铁方向，使每个小磁铁都可以用来储存信息。盘片在电机的带动下高速旋转，日常使用的笔记本硬盘每分钟转速达到5400转，我们的存储设备中使用的高性能硬盘转速每分钟达到7200转、10000转、15000转。转速的不同造成硬盘读写能力的不同，后面我们详细计算他们的IO处理能力。

硬盘的最重要的单元磁头：硬盘的磁头用来读取或者修改盘片上磁性物质的状态，一般来说，每一个磁面都会有一个磁头，从最上面开始，从0开始编号。磁头在停止工作时，与硬盘是接触的，但是在工作时呈飞行状态。磁头采取在盘片的着陆区接触式启停的方式，着陆区不存放任何数据，磁头在此区域启停，不存在损伤任何数据的问题。读取数据时，盘片高速旋转，由于对磁头运动采取了精巧的空气动力学设计，此时磁头处于离盘面数据区0.2---0.5微米高度的"飞行状态"。既不与盘面接触造成磨损，又能读写数据。

电机：硬盘内的电机都为无刷电机，在高速轴承支撑下机械磨损很小，可以长时间连续工作。高速旋转的盘体产生了明显的陀螺效应，所以工作中的硬盘不宜运动，否则将加重轴承的工作负荷。硬盘磁头的寻道饲服电机多采用音圈式旋转或者直线运动步进电机，在饲服跟踪的调节下精确地跟踪盘片的磁道，所以在硬盘工作时不要有冲击碰撞，搬动时要小心轻放。寻道电机控制下的磁头的运动，是左右来回移动的，而且幅度很小，从盘片的最内层（着陆区）启动，慢慢移动到最外层，再慢慢移动回来，一个磁道再到另一个磁道来寻找数据。

硬盘控制器即硬盘控制单元。是把计算机指令转化为硬盘动作的接口设备。它接收并解释计算机来的命令，向硬盘发出各种控制信号。检测硬盘状态，按照规定的硬盘数据格式，把数据写入硬盘和从硬盘读出数据。硬盘控制器类型很多，但它的基本组成和工作原理大体上是相同的，它主要由与计算机系统总线相连的控制逻辑电路，微处理器，完成读出数据分离和写入数据补偿的读写数据解码和编码电路，数据检错和纠错电路，根据计算机发来的命令对数据传递，串并转换以及格式化等进行控制的逻辑电路，存放硬盘基本输入输出程序的只读存储器和用以数据交换的缓冲区等部分组成。

了解了硬盘的构造我们来谈谈对硬盘性能最有影响的指标：IOPS，IOPS (Input/output Per Second)即每秒的读写次数，是衡量磁盘性能的主要指标之一。影响IOPS的因素是IO的服务响应时间，这其中包括一下三项时间。

1. 盘片旋转延迟时间 （rotational latency）
2. 磁头寻道时间 （seek time）
3. 数据传输时间（Data transfer ）

旋转延迟时间，即盘片转到磁头所在位置的时间，由于磁头需要读写的盘片位置是随机的，最远的需要旋转一圈，最近的可能就在磁头所在位置，所以我们取平均值，即盘片旋转半圈的时间来计算，假设硬盘电机转速10000rpm，那旋转一圈的时间是，1/10000=0.0001分钟，换算为毫秒0.0001X60X1000=6毫秒，半圈需要3毫秒。同样可算出15000rpm硬盘延迟时间2毫秒，7200rpm硬盘延迟时间4.17毫秒。

寻道时间：即磁头在盘片径向移动到正确磁道的时间，硬盘厂家标称平均值范围3-15ms，查询到4GB 15000rpm FC硬盘寻道延迟时间约为3.5毫秒，SATA硬盘约为8-9毫秒。

数据传输时间是指完成传输所请求的数据所需要的时间，它取决于数据传输率，其值等于数据大小除以数据传输率。目前IDE/ATA能达到133MB/s，SATA II可达到300MB/s的接口数据传输率，数据传输时间通常远小于前两部分时间。因此，理论上可以忽略。

在日常存储设备选型中，硬盘的物理容量大小及接口类型对单个硬盘IOPS性能的影响可忽略，我们根据不同的硬盘转速及寻道时间计算IO性能来选择，以下常用的硬盘IOps理论值供参考

15000rpm 硬盘 1000/(2+3.5)≈180

10000rpm 硬盘 1000/(3+3.5) ≈150

7200rpm 硬盘 1000/(4.17+8) ≈ 80

IOPS的测试工具IOMETER（http://sourceforge.net/projects/iometer/），可以用于综合测试硬盘的IOPS，由于硬盘的负载类型又分为顺序读写和随机读写，所以客户需要根据应用环境的负载特征，选择合理的应用指标进行测试对比分析，据此选择合适的磁盘类型。