存储原理

对于理解存储，首先从宏观的角度看由应用层到底层分为文件系统-卷管理系统-存储盘控制器-存储盘，今天简单介绍下存储盘。

目前主要存储分为机械磁盘和固态存储SSD，磁盘的机械特性就不再细说，主要介绍一下数据存储的盘片，有盘面、磁道、扇区这几个概念，对应数据存储时的标识就是CHS（cylinder柱面号，header磁头号，sector扇区号），当然磁盘控制器做IO请求时并不是用CHS，而是LBA的编址方式，两者的对应关系表磁盘控制器会保存。Cylinder柱面也是对应磁道，一般限制在1023以下，sector扇区限制在63以下，一个扇区时磁盘的最小单位即512B。此处影响数据处理速度有2个因素：1.交叉因子，它主要是解决由于磁盘转速过快时磁头读取两个或几个连续的扇区时反应不及时导致数据没有读取到，要等再转一圈无形中加长了数据存取时间；2.磁头换道时间较长，而且频繁来回换道也会增加读取时间，在此有设计多种换道方式队列技术，无序传输技术等。

磁盘还有一些缓存的技术例如将磁盘控制器的IO指令缓存优化处理、将取到的数据缓存一次返回就不详细说明，总结来说决定磁盘的性能有3个因素：1.IOPS即每秒磁盘执行IO的次数，2.传输宽带即吞吐量，每次执行IO磁盘能传输的数据量，3.时延即上段描述的时间问题。

现在介绍一下SSD固态存储，它有个显著特点就是相对于机械磁盘来说没有寻道时间，任何地方的开销都是一致的。常见的是基于flash芯片，SSD最小的访问单元时page（4kB），再往上是block，区域。SSD有ECC校验，数据在存取时都需要校验后与校验数据一起放入存储空间。SSD有比较大的一片RAM，通常有64MB甚至有128MB，具体写数据的流程如下：首先假设无free space,先将至少一个block的数据保存到RAM中，将待写入的数据写入到RAM对应位置中，原block全部page放电，然后更新RAM中的数据到真实block上，这有一个概念叫写惩罚。这样做的原因是由于一个block直接写入时会有充电和放电动作会产生干扰导致数据错误，同时page操作粒度太小。前面还提到一个概念是free space，如果存储空间全是free space那写起来就方便多了，因此SSD有记录一份Bitmap，保存了空闲block的信息，备注一点：删除文件只是删除原数据，并不会产生新的free space所以SSD的可用free space会随着使用越来越少。从以上可以看出需要处理好free space防止频繁写惩罚来提高效率和SSD的寿命，例如TRIM可以让文件系统在删除一个文件时及时通知SSD控制器回收对应空间，重写或修改数据时先在free space上写，将原block标记为Garbge，大批量Garbge统一再回收等等。

上面主要介绍了两种硬盘的存储原理，下面说一下磁盘阵列，存储肯定需要大量的硬盘来支撑这就涉及到RAID技术我们逐个解释：1.RAID0,将多个磁盘化成条状，条均匀分布在每个磁盘上，存储时对于大数据如果是一个条的大小刚好分割存储；2.RAID1有备份技术，数据会多备份一次，防止数据丢失；3.RAID2使用校验盘来处理数据正确性，数据不能恢复；4.RAID3改变校验方式XOR，可以恢复丢失数据；5.RAID4如果吞吐量很小时动用所有磁盘去写很耗资源，完全可以写到一个磁盘而其他磁盘用做其他IO，这有个问题就是校验盘每次只能一个IO占用；6.RAID5将校验盘打散分布在每个磁盘上，这样在磁盘数>4时就可以实现并发IO；7.RAID5E/RAID5EE在之前的基础上为了保证磁盘在损坏时能立即备份，需要准备一个备热盘，它不参与组阵，只有在磁盘损坏时才将恢复的数据存在其上，5E与5EE的区别是后者将此备热盘也打散分布在每个磁盘上；8RAID6以上的校验方式都只能恢复一个磁盘数据损坏，当数据盘越来越多有两个同时损坏时就恢复不来，在此基础上实现RAID6，它有两种校验方式可以同时恢复两个磁盘。

附上简单的原理图：

