硬盘IO性能估算入门转载

# ****一、硬盘IO的类型****

IO是输入输出指令，操作系统向存储控制器下发一个读或者写数据的操作指令，控制器下发地址和数据给存储设备，并返回结果给存储控制器，最后到达操作系统。操作系统的一个IO可能会产生多个实际的存储设备IO。一般可以分为：

**1、连续Sequential IO、随机Random IO** 如果本次IO给出的初始地址和上次IO的结束地址是连续或者接近的，磁头可以很快访问到数据，这样的多个IO就是连续IO。如果前后两次IO操作的地址相差比较大，磁头需要较长距离的移动，就是随机IO。在做连续 IO的时候，磁头几乎不换道连接工作，可以快速读取和写入大量数据。

**2、顺序Queue IO、并发Burst IO** 如果存储设备每次只执行一个IO指令，就是顺序IO；当存储设备能同时执行多个IO命令时，控制器发送IO指令后不等回应就继续下发IO指令，称为并发IO。普通磁盘一次只能响应一个IO指令，但RAID等存储设备的IO操作是并发的。对顺序IO的设备，调整IO队列深度，性能不能产生变化。

## ****二、硬盘性能衡量****

衡量硬盘性能，最直观的就是IOPS和吞吐量。

**1、IOPS，每秒处理的IO次数** 指存储设备（HDD、SSD、SAN）单位时间内能处理的IO请求数量，对随机读写频繁的应用，如OLTP数据库、图片、信息，是最关键的衡量指标。IOPS数值受读写比例、随机IO、IO大小、队列深度等因数影响。数据库通常访问一系列不连续的数据，根据文件物理位置，需要很多次IO才能完成。所以需要随机IO高的设备。

**2、Throughput，吞吐量** 指单位时间内可以成功传输的数据数量，传输包括读和写的总和。对于大文件或者流媒体的应用，拥有大量顺序读写，则更关注数据吞吐量。数据吞吐量还受到存储设备接口速度限制，比如IDE、SATA、SAS、FC，SSD使用PCIE Nvme接口最佳，但同时期推出的接口都大于存储设备吞吐量上限。

# ****三、硬盘监控指标****

通过iostat等工具观察存储设备状况，我们一般关注下面指标。

**1、r/s、w/s，每秒读写IO次数。** IOPS = r/s + w/s。读和写的能力，在不同设备中实际是有差异的，但是我们经常简化差异，混合称为IOPS。

**2、avgrq-sz，每次IO请求的平均大小** 如果平均小于32K认为随机存取为主，大于32K则是顺序存储为主。

**3、avgqu-sz，等待执行的IO队列长度** 　IO请求数超出存储设备处理能力，待执行队列就会加长，该值应该小于2.

**4、await，等待执行的IO耗时** 　等待耗时取决svctm以及IO队列长度，该值应该接近svctm。

**5、 svctm，IO服务时间** 处理IO请求的耗时，比如机械硬盘耗时，由寻道时间、旋转延迟和传输时间组成。

**6、uitl，存储设备IO利用率** 存储设备的资源利用情况，队列争用程度成越高值越大。

# ****四、存储性能****

**1、机械硬盘的性能计算** 对机械硬盘来说，IOPS = 1000 / (寻道时间 + 旋转延时 + 传输时间)。 a、寻道时间，磁头移动到对应磁道上的耗时。 Sata 7200RPM，平均寻道时间是9ms。 Sas 10000RPM，平均寻道时间是6ms。 Sas 15000RPM，平均寻道时间是4ms。 　b、旋转延时，盘片旋转至需要的扇区移至磁头下方耗时。 平均为磁盘旋转一周所需时间的一半，60 \* 1000/7200/2。 Sata 7200RPM，旋转延时 60 \* 1000 / 7200 / 2 = 4.17ms。 Sas 10000RPM，旋转延时3ms。 Sas 15000RPM，旋转延时约2ms。 c、传输时间，传输读写数据耗时。 平均数据大小除以接口传输率，耗时很小粗略计算可以忽略。 Sata，300~600MB/s。 Sas，3Gbit/s。 FC，2~4Gbit/s。 NVME，32Gbit/s。 根据上述信息，我们常用的Sas 15000RPM，[MySQL](https://cloud.tencent.com/product/cdb?from=10680" \t "_blank)应用16k块，机械硬盘IOPS = 1000 / (4 + 2 + 16K / 375K) = 165。

**2. SSD硬盘IO计算** 固态硬盘没有寻道时间和旋转时间。IO耗时是通过地址查找数据耗时，根据芯片颗粒SLC、MLC，中控芯片、队列深度32~64、接口Sata、PCIE的不同，一般负载非太高时是相对固定值（控制在60%利用率）。 IOPS = 1000 / IO耗时。因为SSD比较固定，比如Intel 320 SSD对8K avgrq-sz耗时0.1ms，1000/0.1ms=10000 IOPS。具体参考后续文章。

**3、iops估算raid盘数** a、如果单盘不能满足性能需求，需要通过Raid实现，计算所需硬盘数公式如下。 Raid1、10，Drive IOPS = Read IOPS + 2\*Write IOPS Raid3、5，Drive IOPS = Read IOPS + 4\*Write IOPS Raid6，Drive IOPS = Read IOPS + 6\*Write IOPS b、设需要IOPS 5000，读写比例2:1，则RAID10后IOPS需求为。 RAID10，(2/3) \* 5000 + 2 \* (1/3) \* 5000 = 6666 IOPS c、已知Sas 15000RPM机械硬盘IOPS165。 5000IOPS所需硬盘个数, 6666 / 165 = 40。

**4、常见存储设备参考性能，avgrq-sz 8~16K** 5400 rpm SATA，60 IOPS 7200 rpm SATA，70 IOPS 10000 rpm SAS，110 IOPS 15000 rpm SAS，150 IOPS，Sequential RW 180MB/s、Radom RW 15MB/s。 10000 rpm FC，125 IOPS 15000 rpm FC，150 IOPS SSD Sata，3000~40000 IOPS，R 400MB/s、W 250MB/s。 SSD PCIE，20000~40000 IOPS，R 500MB/s、W 300MB/s。 内存，1000000+ IOPS，30~60 GB/s。