

25 | 基础篇：Linux 磁盘I/O是怎么工作的（下）

2019-01-16 倪朋飞

Linux性能优化实战

[进入课程 >](#)



讲述：冯永吉

时长 07:21 大小 6.75M



你好，我是倪朋飞。

上一节我们学习了 Linux 磁盘 I/O 的工作原理，并了解了由文件系统层、通用块层和设备层构成的 Linux 存储系统 I/O 栈。

其中，通用块层是 Linux 磁盘 I/O 的核心。向上，它为文件系统和应用程序，提供访问了块设备的标准接口；向下，把各种异构的磁盘设备，抽象为统一的块设备，并会对文件系统和应用程序发来的 I/O 请求，进行重新排序、请求合并等，提高了磁盘访问的效率。

掌握了磁盘 I/O 的工作原理，你估计迫不及待想知道，怎么才能衡量磁盘的 I/O 性能。

接下来，我们就来看看，磁盘的性能指标，以及观测这些指标的方法。

磁盘性能指标

说到磁盘性能的衡量标准，必须要提到五个常见指标，也就是我们经常用到的，使用率、饱和度和 IOPS、吞吐量以及响应时间等。这五个指标，是衡量磁盘性能的基本指标。

使用率，是指磁盘处理 I/O 的时间百分比。过高的使用率（比如超过 80%），通常意味着磁盘 I/O 存在性能瓶颈。

饱和度，是指磁盘处理 I/O 的繁忙程度。过高的饱和度，意味着磁盘存在严重的性能瓶颈。当饱和度为 100% 时，磁盘无法接受新的 I/O 请求。

IOPS（Input/Output Per Second），是指每秒的 I/O 请求数。

吞吐量，是指每秒的 I/O 请求大小。

响应时间，是指 I/O 请求从发出到收到响应的间隔时间。

这里要注意的是，使用率只考虑有没有 I/O，而不考虑 I/O 的大小。换句话说，当使用率是 100% 的时候，磁盘依然有可能接受新的 I/O 请求。

这些指标，很可能是你经常挂在嘴边的，一讨论磁盘性能必定提起的对象。不过我还是要强调一点，不要孤立地去比较某一指标，而要结合读写比例、I/O 类型（随机还是连续）以及 I/O 的大小，综合来分析。

举个例子，在数据库、大量小文件等这类随机读写比较多的场景中，IOPS 更能反映系统的整体性能；而在多媒体等顺序读写较多的场景中，吞吐量才更能反映系统的整体性能。

一般来说，我们在为应用程序的服务器选型时，要先对磁盘的 I/O 性能进行基准测试，以便可以准确评估，磁盘性能是否可以满足应用程序的需求。

这一方面，我推荐用性能测试工具 fio，来测试磁盘的 IOPS、吞吐量以及响应时间等核心指标。但还是那句话，因地制宜，灵活选取。在基准测试时，一定要注意根据应用程序 I/O 的特点，来具体评估指标。

当然，这就需要你测试出，不同 I/O 大小（一般是 512B 至 1MB 中间的若干值）分别在随机读、顺序读、随机写、顺序写等各种场景下的性能情况。

用性能工具得到的这些指标，可以作为后续分析应用程序性能的依据。一旦发生性能问题，你就可以把它们作为磁盘性能的极限值，进而评估磁盘 I/O 的使用情况。

了解磁盘的性能指标，只是我们 I/O 性能测试的第一步。接下来，又该用什么方法来观测它们呢？这里，我给你介绍几个常用的 I/O 性能观测方法。

磁盘 I/O 观测

第一个要观测的，是每块磁盘的使用情况。

iostat 是最常用的磁盘 I/O 性能观测工具，它提供了每个磁盘的使用率、IOPS、吞吐量等各种常见的性能指标，当然，这些指标实际上来自 /proc/diskstats。

iostat 的输出界面如下。

复制代码

```
1 # -d -x 表示显示所有磁盘 I/O 的指标
2 $ iostat -d -x 1
3 Device      r/s      w/s      rkB/s      wkB/s      rrqm/s      wrqm/s      %rrqm      %wrqm      r_await
4 loop0        0.00      0.00        0.00        0.00        0.00        0.00        0.00        0.00        0.00
5 loop1        0.00      0.00        0.00        0.00        0.00        0.00        0.00        0.00        0.00
6 sda          0.00      0.00        0.00        0.00        0.00        0.00        0.00        0.00        0.00
7 sdb          0.00      0.00        0.00        0.00        0.00        0.00        0.00        0.00        0.00
```

从这里你可以看到，iostat 提供了非常丰富的性能指标。第一列的 Device 表示磁盘设备的名字，其他各列指标，虽然数量较多，但是每个指标的含义都很重要。为了方便你理解，我把它们总结成了一个表格。

iostat 指标解读		
性能指标	含义	提示
r/s	每秒发送给磁盘的读请求数	合并后的请求数
w/s	每秒发送给磁盘的写请求数	合并后的请求数
rkB/s	每秒从磁盘读取的数据量	单位为kB
wkB/s	每秒向磁盘写入的数据量	单位为kB
rrqm/s	每秒合并的读请求数	%rrqm表示合并读请求的百分比
wrqm/s	每秒合并的写请求数	%wrqm表示合并写请求的百分比
r_await	读请求处理完成等待时间	包括队列中的等待时间和设备实际处理的时间，单位为毫秒
w_await	写请求处理完成等待时间	包括队列中的等待时间和设备实际处理的时间，单位为毫秒
aqu-sz	平均请求队列长度	旧版中为avgqu-sz
rareq-sz	平均读请求大小	单位为kB
wareq-sz	平均写请求大小	单位为kB
svctm	处理I/O请求所需的平均时间（不包括等待时间）	单位为毫秒。注意这是推断的数据，并不保证完全准确
%util	磁盘处理I/O的时间百分比	即使用率，由于可能存在并行I/O，100%并不一定表明磁盘I/O饱和

这些指标中，你要注意：

- %util ，就是我们前面提到的磁盘 I/O 使用率；
- r/s+ w/s ，就是 IOPS ；
- rkB/s+wkB/s ，就是吞吐量；
- r_await+w_await ，就是响应时间。

在观测指标时，也别忘了结合请求的大小（ rareq-sz 和 wareq-sz ）一起分析。


你可能注意到，从 `iostat` 并不能直接得到磁盘饱和度。事实上，饱和度通常也没有其他简单的观测方法，不过，你可以把观测到的，平均请求队列长度或者读写请求完成的等待时间，跟基准测试的结果（比如通过 `fio`）进行对比，综合评估磁盘的饱和情况。

进程 I/O 观测

除了每块磁盘的 I/O 情况，每个进程的 I/O 情况也是我们需要关注的重点。

上面提到的 `iostat` 只提供磁盘整体的 I/O 性能数据，缺点在于，并不能知道具体是哪些进程在进行磁盘读写。要观察进程的 I/O 情况，你还可以使用 `pidstat` 和 `iotop` 这两个工具。

`pidstat` 是我们的老朋友了，这里我就不再啰嗦它的功能了。给它加上 `-d` 参数，你就可以看到进程的 I/O 情况，如下所示：

 复制代码

```
1 $ pidstat -d 1
2 13:39:51      UID      PID   kB_rd/s   kB_wr/s kB_ccwr/s iodelay  Command
3 13:39:52      102     916     0.00     4.00     0.00      0  rsyslogd
```

从 `pidstat` 的输出你能看到，它可以实时查看每个进程的 I/O 情况，包括下面这些内容。

用户 ID（UID）和进程 ID（PID）。

每秒读取的数据大小（kB_rd/s），单位是 KB。


每秒发出的写请求数据大小（kB_wr/s），单位是 KB。

每秒取消的写请求数据大小（kB_ccwr/s），单位是 KB。

块 I/O 延迟（iodelay），包括等待同步块 I/O 和换入块 I/O 结束的时间，单位是时钟周期。

除了可以用 `pidstat` 实时查看，根据 I/O 大小对进程排序，也是性能分析中一个常用的方法。这一点，我推荐另一个工具，`iotop`。它是一个类似于 `top` 的工具，你可以按照 I/O 大小对进程排序，然后找到 I/O 较大的那些进程。

iotop 的输出如下所示：

 复制代码

```
1 $ iotop
2 Total DISK READ :      0.00 B/s | Total DISK WRITE :      7.85 K/s
3 Actual DISK READ:      0.00 B/s | Actual DISK WRITE:      0.00 B/s
4   TID  PRIO  USER      DISK READ  DISK WRITE  SWAPIN      IO>     COMMAND
5 15055 be/3 root          0.00 B/s    7.85 K/s   0.00 %   0.00 % systemd-journald
```

从这个输出，你可以看到，前两行分别表示，进程的磁盘读写大小总数和磁盘真实的读写大小总数。因为缓存、缓冲区、I/O 合并等因素的影响，它们可能并不相等。

剩下的部分，则是从各个角度来分别表示进程的 I/O 情况，包括线程 ID、I/O 优先级、每秒读磁盘的大小、每秒写磁盘的大小、换入和等待 I/O 的时钟百分比等。

这两个工具，是我们分析磁盘 I/O 性能时最常用到的。你先了解它们的功能和指标含义，具体的使用方法，接下来的案例实战中我们一起学习。

小结

今天，我们梳理了 Linux 磁盘 I/O 的性能指标和性能工具。我们通常用 IOPS、吞吐量、使用率、饱和度以及响应时间等几个指标，来评估磁盘的 I/O 性能。

你可以用 iostat 获得磁盘的 I/O 情况，也可以用 pidstat、iotop 等观察进程的 I/O 情况。不过在分析这些性能指标时，你要注意结合读写比例、I/O 类型以及 I/O 大小等，进行综合分析。

思考

最后，我想请你一起来聊聊，你碰到过的磁盘 I/O 问题。在碰到磁盘 I/O 性能问题时，你是怎么分析和定位的呢？你可以结合今天学到的磁盘 I/O 指标和工具，以及上一节学过的磁盘 I/O 原理，来总结你的思路。

欢迎在留言区和我讨论，也欢迎把这篇文章分享给你的同事、朋友。我们一起在实战中演练，在交流中进步。

Linux 性能优化实战

10 分钟帮你找到系统瓶颈

倪朋飞

微软资深工程师
Kubernetes 项目维护者



新版升级：点击「 请朋友读」，10位好友免费读，邀请订阅更有**现金**奖励。

© 版权归极客邦科技所有，未经许可不得传播售卖。页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 24 | 基础篇：Linux 磁盘I/O是怎么工作的（上）

下一篇 26 | 案例篇：如何找出狂打日志的“内鬼”？

精选留言 (24)

 写留言



每天晒白牙

2019-01-17

 6

【D25打卡】

总结：

磁盘性能检测指标：

使用率：磁盘处理I/O的时间百分比，使用率只考虑有没有I/O，不考虑I/O的大小。注意当使用率为100%时，由于可能存在并行I/O，磁盘并不一定饱和，所以磁盘仍然可能接收...

展开 ▾

作者回复: 使用率是从时间角度衡量I/O，但是磁盘还可以支持并行写，所以即使使用率100%，有可能还可以接收新的I/O（不饱和）



ninuxer

2019-01-16

👍 3

day26打卡

之前都没用过fio测试磁盘实际性能，基本都是依赖磁盘型号查官网数据作为依据~
iostat和iotop倒是会经常用，之前有几列输出的内容自己理解有偏差，这下算是纠正过来了👉

展开 ▾



Cranliu

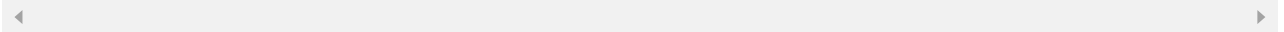
2019-01-16

👍 3

关于磁盘的饱和度，有没有经验值可以参考下呢？谢谢

展开 ▾

作者回复: 饱和度一般没法直接观测到，所以一般是通过实际观测值跟基准测试结果对比来分析



Ender0224

2019-02-10

👍 2

仲鬼

2019-01-25

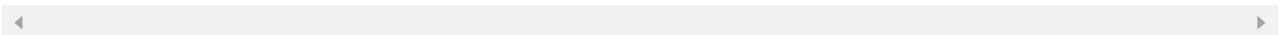
□

2

"r_await+w_await，就是响应时间"...

展开 ▾

作者回复: 是的



仲鬼

2019-01-25

👍 2

"r_await+w_await，就是响应时间"

对这句表述有怀疑。

r_await、w_await分别是读、写请求的平均等待时间，二者相加什么都不是。因为 $a/b + c/d \neq (a+c)/(b+d)$ 。

展开 □...

展开 ▾



仲鬼

2019-01-18

👍 2

" $r_await + w_await$ ，就是响应时间"

对这句表述有怀疑。

r_await 、 w_await 分别是读、写请求的平均等待时间，二者相加什么都不是。因为 $a/b + c/d$ 不等于 $(a+c)/(b+d)$ 。

展开 ▾

作者回复: 从公式上是这样，但间隔时间相同的时候呢？



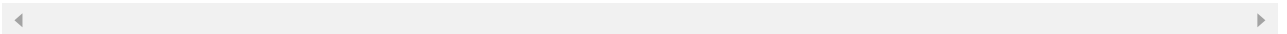
Christmas

2019-01-16

👍 2

一趟调度法，电梯调度法等调度是发生在磁盘控制器硬件上的吗？通用块层的调度是os级别的对吧？

作者回复: 是的



remcarpedi...

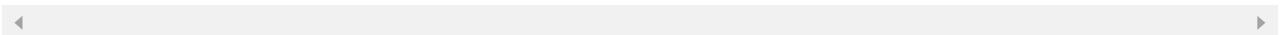
2019-01-21

👍 1

请问作者对《性能之垫-洞悉系统、企业和云计算》这本书的看法？适合作为工具书，用于查阅；还是可以进行通篇学习

展开 ▾

作者回复: 建议学习一下各个章节的基本原理和思路，剩下的工具部分作为手册参考。不过有些工具过时了，使用的时候要注意



Boy-stru...

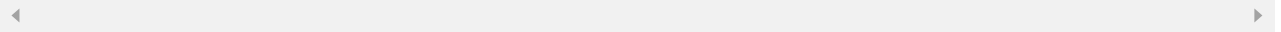
2019-04-03

👍

老师，如何根据系统调用判断IO为随机还是顺序，IO 的位置怎么体现，希望老师可以结合案例具体讲解一下，多谢！

展开 ▾

作者回复: 最简单的方法是根据系统调用判断I/O读写的相对位置



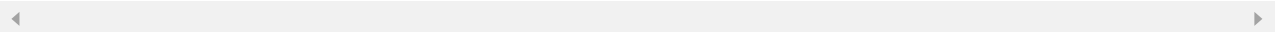
张挺

2019-03-29



使用率指标不太理解，请问这个值是怎么计算出来的呢？

作者回复: 使用率，是指磁盘处理 I/O 的时间百分比



Vincent

2019-03-24



随机io和顺序io就跟数据结构有关系了吧？比如数组和链表。除了通过代码判断是随机io还是顺序io 系统有什么工具可以判断吗？

展开 ∨

作者回复: 数组和链表还是内存中的数据结构，I/O是指跟磁盘的交互。跟踪进程的系统调用或者磁盘的I/O，根据读写的相对位置可以判断



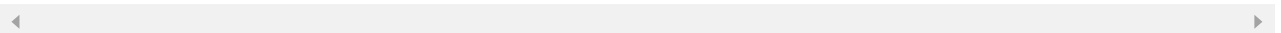
挺直腰板

2019-03-24



老师，如何知道是随机IO还是顺序IO,两者性能差还是蛮大

作者回复: 可以通过系统调用观察I/O的相对位置



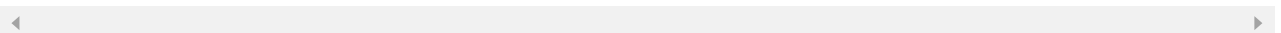
苦行僧

2019-03-04



老师在使用网络挂载的共享存储io性能差，有什么优化方式吗

作者回复: 这要看场景的，需要先定位出瓶颈是哪里导致的





jssfy

2019-01-24



iotop可以看到在nfs上的流量不？

展开 ∨

作者回复: 应该是可以的，不过我没有试过，你可以执行一下看看



刘涛^_^

2019-01-22



老师，IO的饱和度怎么衡量

展开 ∨



dexter

2019-01-18



打卡

展开 ∨



fran712

2019-01-17



请问将/dev/sda直接挂载到某个目录和将磁盘只一个分区后，/dev/sda1挂载到某个目录，这两种挂载的区别是什么？

展开 ∨

作者回复: 区别是你怎么创建文件系统的，是使用分区还是整块磁盘



一生一世

2019-01-17



老师能否提供一些参数性能指标参考，有时候能看到指标却无法确定是否有问题

作者回复: 这需要基准测试的，我的机器指标很可能不适合你的环境



行行行

2019-01-17



老师页框回收算法可以讲下吗

展开 ▼



我来也

2019-01-16



[D25打卡]

今天又见到了新工具FIO和iotop

之前都是用的 vmstat iostat pidstat .

以前没有这么精细的分析i/o.因为程序的瓶颈不在这块.