



30 | 套路篇：如何迅速分析出系统I/O的瓶颈在哪里？


導讀 by 泳褲





文件系统 I/O 性能指标

- 存儲空間 (容量、使用量以及剩餘空間)
 - 經過文件系統向外展示的數字
 - metadata也會占用空間
 - 不代表真實使用的硬碟大小
 - 如果有用RAID會差更多
 - inode (容量、使用量以及剩餘量)
 - 初學者常沒注意到的部分

- 
- 緩存使用情況
 - 页缓存 page cache
 - 目录项缓存 dentry cache
 - 索引节点缓存 Inode Cache
 - 各文件系统緩存 (ext4、XFS等)



磁碟IO性能指標

- 使用率
 - 磁碟處理IO的時間百分比 如果使用率太高 通常意味著磁碟IO有效能瓶頸
 - 使用率不考慮IO的大小
- 飽和度
 - 磁碟處理IO的繁忙程度 過高的飽和度 意味著磁碟有嚴重的效能瓶頸
 - 當飽和度100%時 磁盤無法接受新的IO請求
- IOPS
 - 每秒的IO請求數
- 吞吐量
 - 每秒的IO請求大小
- 響應時間



注意综合 I/O 的具体场景

- 讀寫類型(循序或隨機)
- 讀寫比例
- 讀寫大小
- 儲存類型
 - 有無RAID
 - RAID類型
 - 本地
 - 網路
 - 是不是SSD

**不同場景的IO性能指標
不能直接分析比對**

—

缓冲区(Buffer)也是要重点掌握的指标, 它经常出现在内存和磁盘问题的分析中

—



性能工具

—



文件系統

- 可以看空間容量也能看inode的使用狀況
 - df
- cache
 - /proc/meminfo
 - /proc/slabinfo
 - slabtop



磁碟IO

- iostat
- pidstat

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

```
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。



狂打日誌的案例

- top看CPU 發現iowait比較高
- iostat發現硬碟的IO使用率瓶頸
- pidstat找出大量IO的process
- strace和lsof找出process正在讀寫的文件
- 最後鎖定問題



磁碟IO延遲

- top看CPU 發現iowait比較高
- iostat發現硬碟的IO使用率瓶頸
- pidstat找出process
- 但是strace找不到
- 改用filetop跟opensnoop 從內核找出瓶頸



Mysql

- top看CPU 發現iowait比較高
- iostat發現硬碟的IO使用率瓶頸
- pidstat找出process
- strace+lsof找出正在讀寫的檔案
- 使用mysql指令找出正在運行的工作
- 找出query慢的原因 (index建立失敗)
- 解決index建立失敗的問題建立好index以後問題解決



Redis

- top看CPU 發現iowait比較高
- iostat發現硬碟的IO使用率瓶頸
- pidstat找出process
- strace+lsof找出正在讀寫的檔案
- 從redis工作原理推斷式持久化設定有問題(appendfsync 改成 everysec)
- 修正設定以後發現依然有部份問題
- 從程式面修正行為(塞自己記憶體就好)

性能指標和工具的聯繫

從文件系統和磁碟io性能指標出發

根据指标找工具（文件系统和磁盘I/O）		
性能指标	工具	说明
文件系统空间容量、使用量以及剩余空间	df	详细文档 见 info coreut ils 'df invocation'
索引节点容量、使用量以及剩余量	df	使用 -i 选项
页缓存和可回收Slab缓存	/proc/meminfo sar、vmstat	使用 sar -r 选项
缓冲区	/proc/meminfo sar、vmstat	使用 sar -r 选项
目录项、索引节点以及文件系统的缓存	/proc/slabinfo slabtop	slabtop更直观
磁盘 I/O 使用率、IOPS、吞吐量、响应时间、I/O平均大小以及等待队列长度	iostat sar、dstat	使用 iostat -d -x 或 sar -d 选项
进程I/O大小以及I/O延迟	pidstat iotop	使用 pidstat -d 选项
块设备 I/O 事件跟踪	blktrace	示例: blktrace -d /dev/sda -o- blkparse -i-
进程 I/O 系统调用跟踪	strace	通过系统调用跟踪进程的 I/O
进程块设备I/O大小跟踪	biosnoop biotop	需要安装bcc软件包

從工具出發

根据工具查指标（文件系统和磁盘I/O）	
性能工具	性能指标
iostat	磁盘 I/O 使用率、IOPS、吞吐量、响应时间、I/O平均大小以及等待队列长度
pidstat	进程 I/O 大小以及 I/O 延迟
sar	磁盘 I/O 使用率、IOPS、吞吐量以及响应时间
dstat	磁盘 I/O 使用率、IOPS以及吞吐量
iotop	按 I/O 大小对进程排序
slabtop	目录项、索引节点以及文件系统的缓存
/proc/slabinfo	目录项、索引节点以及文件系统的缓存
/proc/meminfo	页缓存和可回收Slab缓存
/proc/diskstats	磁盘的 IOPS、吞吐量以及延迟
/proc/pid/io	进程IOPS、 I/O 大小以及 I/O 延迟
vmstat	缓存和缓冲区用量汇总
blktrace	跟踪块设备I/O事件
biosnoop	跟踪进程的块设备I/O大小
biotop	跟踪进程块I/O并按I/O大小排序
strace	跟踪进程的I/O系统调用
perf	跟踪内核中的I/O事件
df	磁盘空间和索引节点使用量和剩余量
mount	文件系统的挂载路径以及挂载参数
du	目录占用的磁盘空间大小
tune2fs	显示和设置文件系统参数
hdparam	显示和设置磁盘参数

如何迅速分析io的性能瓶頸



分析思路

- top看CPU 發現iowait比較高
- iostat發現硬碟的IO使用率瓶頸
- pidstat找出process
- 分析行為
- 從process的工作找出來源



從廣到窄

先用指標多的工具

在慢慢收窄範圍

