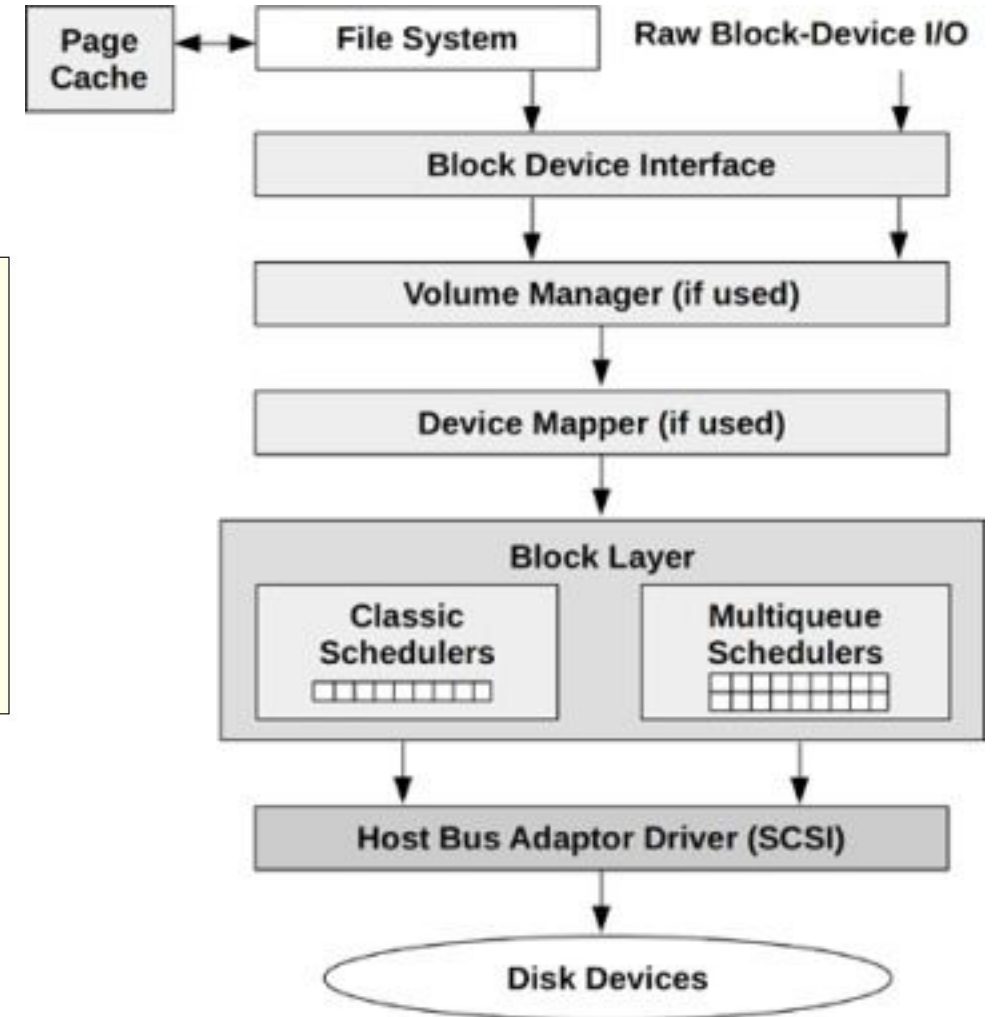
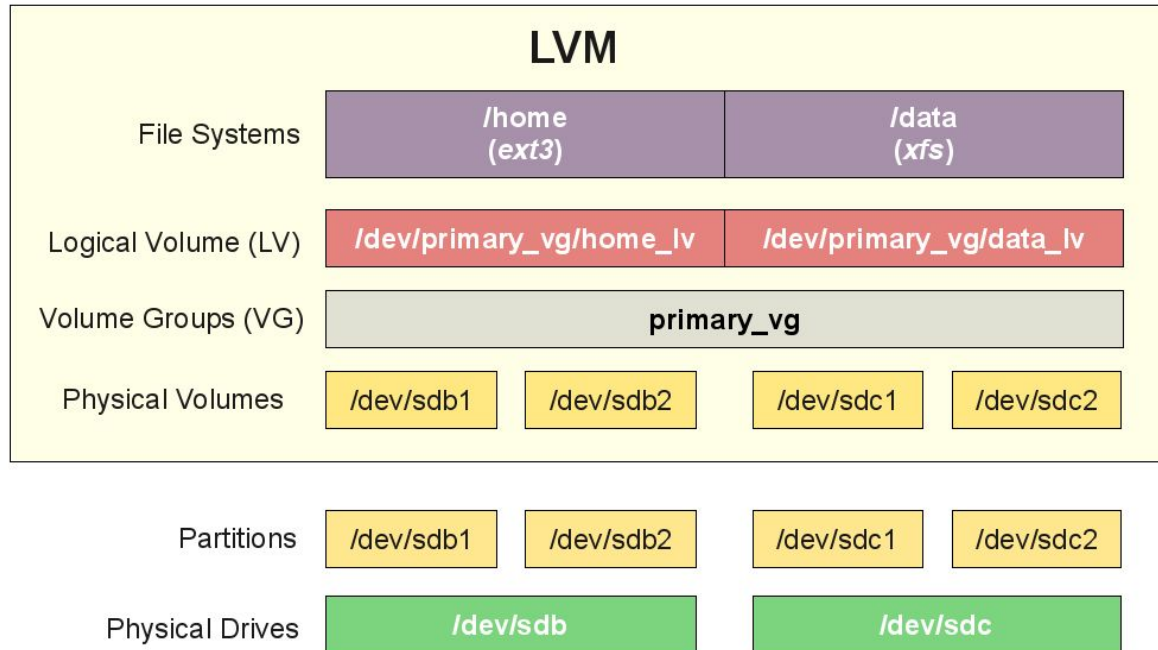




25 | 基礎篇 : Linux 硬碟I/O是 怎麼工作的(下)

Hazel Shen

Linux Block I/O Stack



Source: <https://www.informit.com/articles/article.aspx?p=2995360>

Source:

https://access.redhat.com/documentation/zh-tw/red_hat_enterprise_linux/6/html/logical_volume_manager_administration/lvm_cluster_overview

Source: https://access.redhat.com/documentation/zh-tw/red_hat_enterprise_linux/6/html/logical_volume_manager_administration/device_mapper

- 文件系統層
- 通用層
- 設備層





常見 硬碟性能指標

- 使用率: 處理 I/O 時間
- 飽和度: 處理 I/O 繁忙度
- IOPS: 每秒 I/O 請求數
- Throughput: 每秒 I/O 請求大小 e.g. MB/s

$$\text{Throughput MB/s} = \text{IOPS} * \text{KB per IO} / 1024$$

- Response Time

還需要參考什麼標準？



硬碟讀寫比例

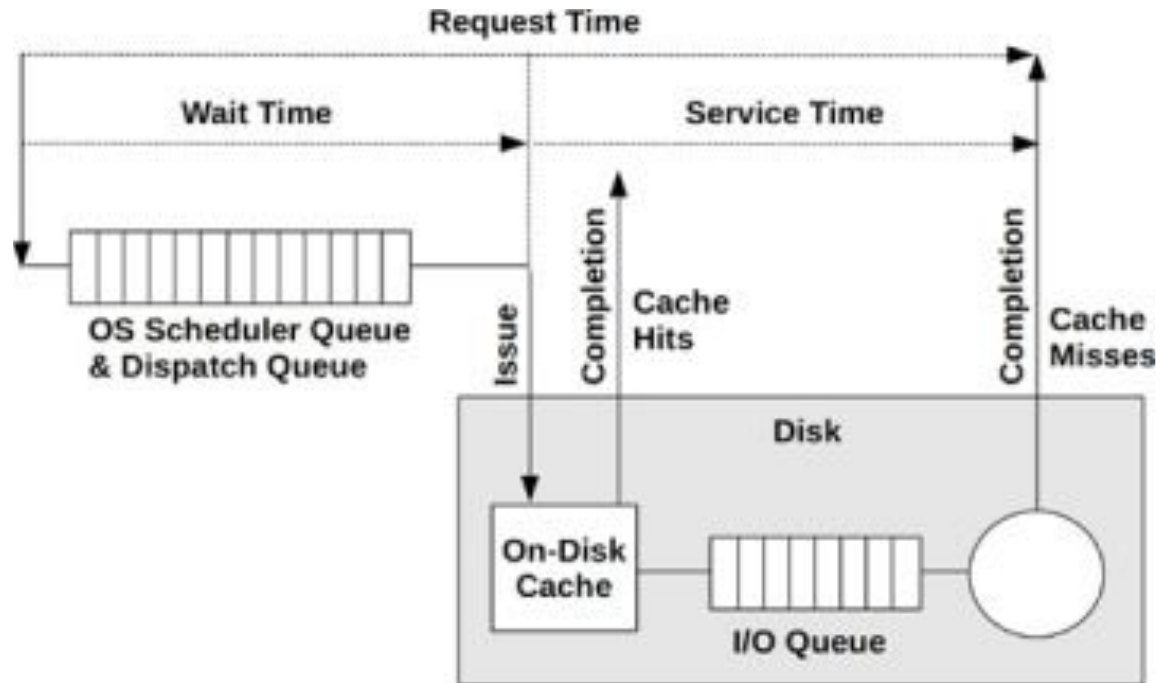


I/O 類型 E.G. 隨機 / 連續




I/O 大小

Disk I/O with OS



為甚麼不能純用使用率判別 I/O 效能？



舉個栗子 – 隨機讀寫

- 資料庫 Databases
- 大量小文件讀寫

| 應用類型 | IO大小 | 讀寫比例 | 隨機與循序讀寫比例 |
|-----------------|-----------------|-----------|-------------|
| Web File Server | 4KB、8KB、64KB | 95%讀/5%寫 | 75%隨機/25%循序 |
| Web Server Log | 8KB | 100% 寫 | 100%循序 |
| OS Paging | 64KB | 90%讀/10%寫 | 100%循序 |
| Exchange Server | 4KB | 67%讀/33%寫 | 100%隨機 |
| Workstation | 8KB | 80%讀/20%寫 | 80%隨機/20%循序 |
| Media Streaming | 64KB | 98%讀/2%寫 | 100%循序 |
| OLTP - Data | 8KB | 70%讀/30%寫 | 100%隨機 |
| OLTP - Log | 512bytes - 64KB | 100%寫 | 100%循序 |

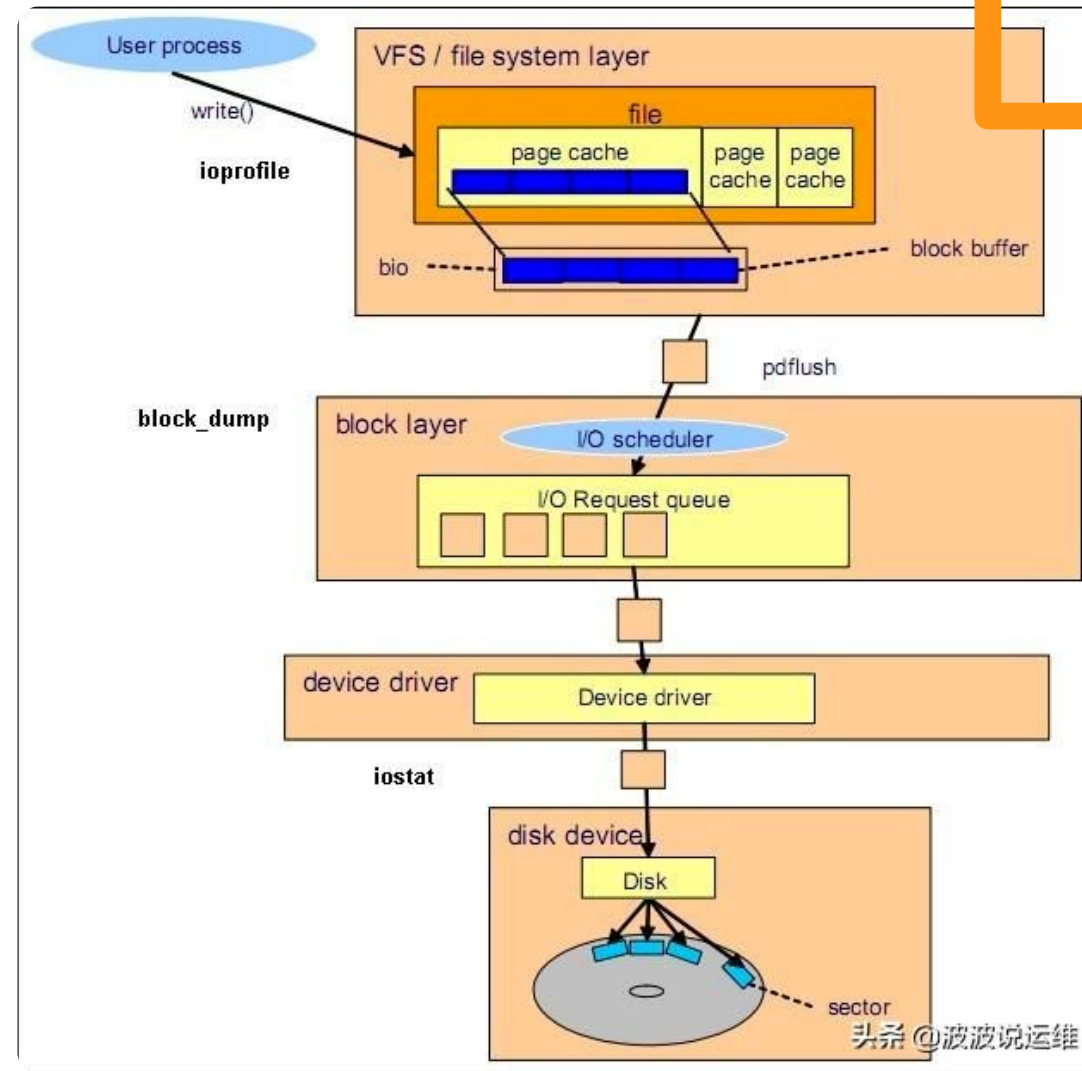
工具與 Linux 結構位置對應

文件級: lsof, ls /proc/pid/fd

APP級: ioprofile (pt-ioprofile)

Process 級: iotop, pidstat

系統級: iostat



硬碟 I/O 觀察(iostat from /proc/diskstats)

-d -x表示所有硬碟/I/O的指標

```
[root@lb ~]# iostat -d -x 1
Linux 3.10.0-1127.13.1.el7.x86_64 (lb.ocp.hazel)      07/19/2020      _x86_64_      (4 CPU)

Device:            rrqm/s    wrqm/s      r/s      w/s    rkB/s    wkB/s avgrq-sz avgqu-sz   await  r_await  w_await  svctm  %util
sda                 0.05      1.63    30.31    11.03  1079.50  2573.31   176.71     2.19   52.96    4.67   185.70    2.08   8.62
dm-0                 0.00      0.00    28.64    12.46  1006.80  2513.29   171.29     2.32   56.51    4.85   175.23    2.06   8.48
dm-1                 0.00      0.00     0.34     0.00     8.56     0.00    50.09     0.00    1.70    1.70     0.00    1.51   0.05
```

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 |

IOPS

Throughput

Response Time

➔ I/O 使用率

Process I/O 觀察(pidstat: 即時查看 Process 消耗資源)

加上 -C 可以篩選 Command

```
[root@lb ~]# pidstat -d
Linux 3.10.0-1127.13.1.el7.x86_64 (lb.ocp.hazel)          07/19/2020      _x86_64_      (4 CPU)

09:24:54 AM    UID      PID    kB_rd/s    kB_wr/s  kB_ccwr/s  Command
09:24:54 AM      0        1     105.56     97.70      36.45    systemd
09:24:54 AM      0       406       0.01       0.00       0.00    kworker/u8:30
09:24:54 AM      0       600       1.06       0.00       0.00    systemd-journal
09:24:54 AM      0       620       0.08       0.00       0.00    lvmetad
09:24:54 AM      0       637      13.04       0.00       0.00    systemd-udev
```

```
[root@lb ~]# pidstat -d -C auditd
Linux 3.10.0-1127.13.1.el7.x86_64 (lb.ocp.hazel)          07/19/2020      _x86_64_

09:25:09 AM    UID      PID    kB_rd/s    kB_wr/s  kB_ccwr/s  Command
09:25:09 AM      0       772       0.00       0.14       0.00    auditd
```

Process I/O 觀察 (iotop)

```
Total DISK READ :      0.00 B/s | Total DISK WRITE :      0.00 B/s
Actual DISK READ:      0.00 B/s | Actual DISK WRITE:      0.00 B/s
```

| TID | PRI | USER | DISK READ | DISK WRITE | SWAPIN | IO> | COMMAND |
|-----|------|------|-----------|------------|--------|--------|---|
| 512 | be/0 | root | 0.00 B/s | 0.00 B/s | 0.00 % | 0.00 % | [xfs-data/dm-0] |
| 1 | be/4 | root | 0.00 B/s | 0.00 B/s | 0.00 % | 0.00 % | systemd --switched-root --system --deserialize 22 |
| 2 | be/4 | root | 0.00 B/s | 0.00 B/s | 0.00 % | 0.00 % | [kthreadd] |
| 515 | be/0 | root | 0.00 B/s | 0.00 B/s | 0.00 % | 0.00 % | [xfs-reclaim/dm-] |
| 4 | be/0 | root | 0.00 B/s | 0.00 B/s | 0.00 % | 0.00 % | [kworker/0:0H] |

Fio 硬碟性能測試工具

測試方法：

相同 I/O 大小 (1MB) 分別

隨機讀

循序讀

隨機寫

循序寫

混和讀寫

Source: https://fio.readthedocs.io/en/latest/fio_doc.html

Source: <http://benjr.tw/34632>

隨機讀 randread

```
fio --name=random-writers --ioengine=libaio --iodepth=4 --rw=randread \  
--bs=1024k --direct=0 --size=64m --numjobs=1
```

Run status group 0 (all jobs):

READ: bw=94.3MiB/s (98.8MB/s), 94.3MiB/s-94.3MiB/s (98.8MB/s-98.8MB/s),
io=64.0MiB (67.1MB), run=679-679msec

Disk stats (read/write):

dm-0: ios=124/0, merge=0/0, ticks=1007/0, in_queue=1023, util=81.82%
, aggrios=128/0, aggrmerge=0/0, aggrticks=1073/0, aggrin_queue=1071, aggrutil=80.00%

sda: ios=128/0, merge=0/0, ticks=1073/0, in_queue=1071, util=80.00%

循序讀 read

```
fio --name=random-writers --ioengine=libaio --iodepth=4 --rw=read \  
--bs=1024k --direct=0 --size=64m --numjobs=1
```

Run status group 0 (all jobs):

READ: bw=162MiB/s (170MB/s), 162MiB/s-162MiB/s (170MB/s-170MB/s),
io=64.0MiB (67.1MB), run=394-394msec

Disk stats (read/write):

dm-0: ios=60/0, merge=0/0, ticks=1822/0, in_queue=2595, util=54.44%
,aggrios=128/0, aggrmerge=0/0, aggrticks=7600/0, aggrin_queue=7599, aggrutil=74.12%

sda: ios=128/0, merge=0/0, ticks=7600/0, in_queue=7599, util=74.12%

隨機寫 randwrite

```
fio --name=random-writers --ioengine=libaio --iodepth=4 --rw=randwrite \  
--bs=1024k --direct=0 --size=64m --numjobs=1
```

Run status group 0 (all jobs):

WRITE: bw=1684MiB/s (1766MB/s), 1684MiB/s-1684MiB/s (1766MB/s-1766MB/s),
io=64.0MiB (67.1MB), run=38-38msec

Disk stats (read/write):

dm-0: ios=0/0, merge=0/0, ticks=0/0, in_queue=0, util=0.00%
, aggrios=0/24, aggrmerge=0/0, aggrticks=0/474, aggrin_queue=5426, aggrutil=34.78%

sda: ios=0/24, merge=0/0, ticks=0/474, in_queue=5426, util=34.78%

循序寫 write

```
fio --name=random-writers --ioengine=libaio --iodepth=4 --rw=write \  
--bs=1024k --direct=0 --size=64m --numjobs=1
```

Run status group 0 (all jobs):

WRITE: bw=1730MiB/s (1814MB/s), 1730MiB/s-1730MiB/s (1814MB/s-1814MB/s)
 , io=64.0MiB (67.1MB), run=37-37msec

Disk stats (read/write):

dm-0: ios=0/0, merge=0/0, ticks=0/0, in_queue=0, util=0.00%
 , aggrios=0/24, aggrmerge=0/0, aggrticks=0/476, aggrin_queue=5463, aggrutil=34.78%

sda: ios=0/24, merge=0/0, ticks=0/476, in_queue=5463, util=34.78%



小結

- 介紹硬碟性能指標
 - IOPS / throughput / utility rate / saturation
- 介紹硬碟觀察工具
 - fio / pidstat / iotop
- 硬碟壓力測試工具
 - iometer
- 評估的考量要結合：
 - 讀寫比 / IO類型 / IO大小

問題思考

1. 使用率、飽和率, 哪個指標更實用?
2. 飽和度如何觀察 – 比方說花費時間
3. 隨機和循序 I/O 怎麼看?



Thank You