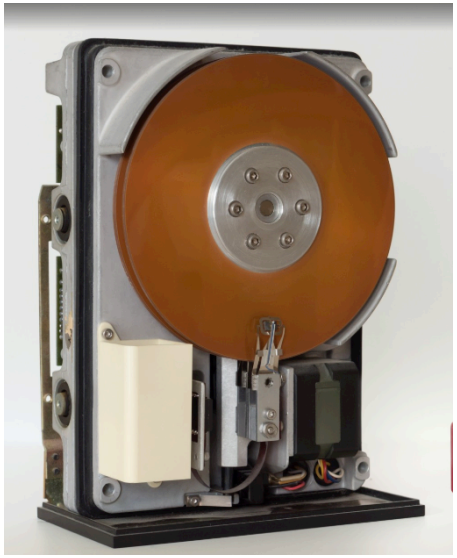
ioTrace
Another Disk Activity Tracing Tool

Ahao Mu
(ahao.mah@alibaba-inc.com)

June 26, 2018

Background

- Requirement proposed by Alibaba's business line: Process centralized disk activities.
- Currently implemented tools can't meet the requirement.



Pain

- The PID/TID are unknown in scenario of disk bandwidth is overhauled.
- It brings difficulties to narrow down the problematic processes/threads.

Disk IO Toolset

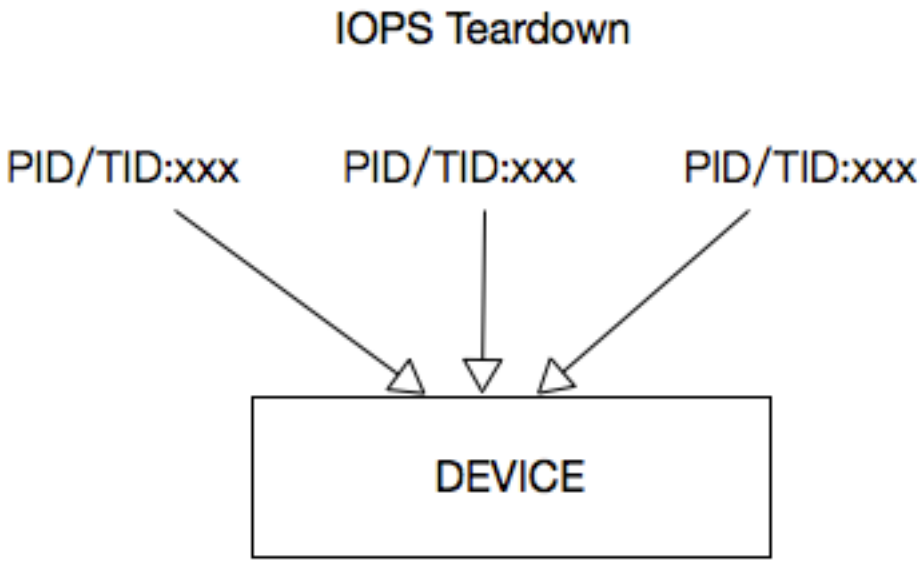
- iotop
  - Written in Python language, read from /proc/<pid>/io and /proc/diskstats.
  - Missed DEVICE dimension.
- iostat
  - Written in C language, read from /proc/diskstats, See Documentation/iostats.txt.
  - Regardless of processes.
- blktrace
  - Written in C language, massive and bogus output.
  - Tremendous performance overhead.

As above all are not the ideal way in our production environment.

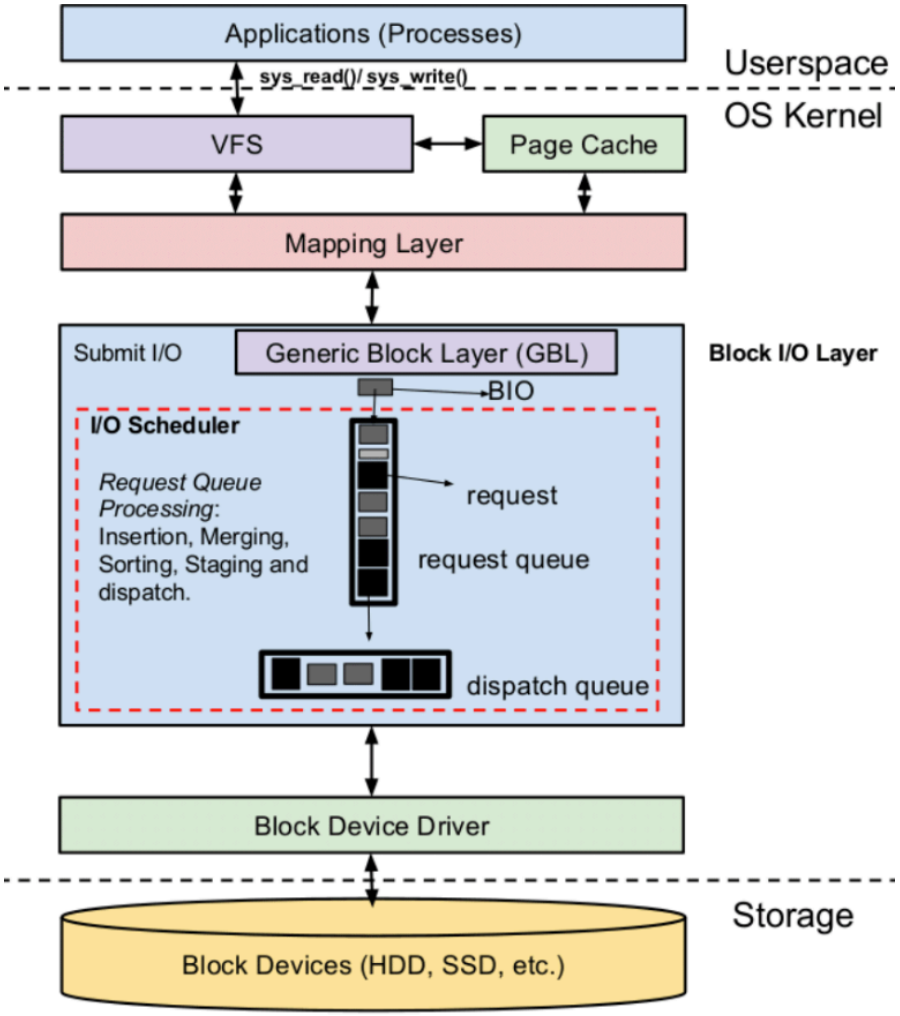
Goal of ioTrace

- Aware of PID/TID and DEVICE dimensions.
- Debugging and monitoring disk's activities.
- Light, agile and easy for daemonizing in production environment.

ioTrace

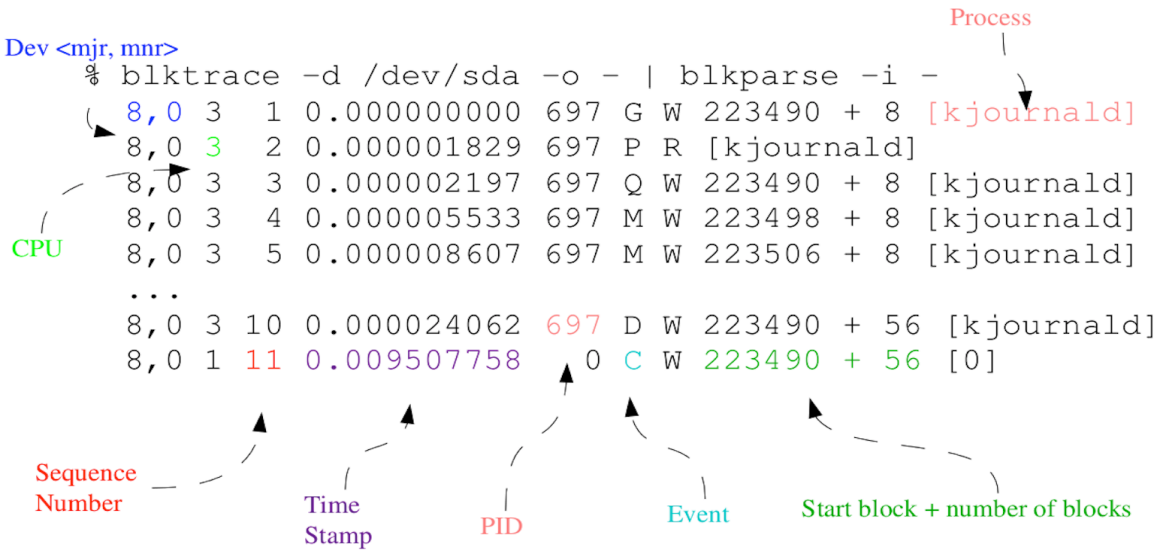


IO Stack


Applications (Processes)
sys_read() / sys_write()
Userspace
OS Kernel
VFS
Page Cache
Mapping Layer
Submit I/O
Generic Block Layer (GBL)
Block I/O Layer
BIO
I/O Scheduler
Request Queue Processing: Insertion, Merging, Sorting, Staging and dispatch.
request
request queue
dispatch queue
Block Device Driver
Storage
Block Devices (HDD, SSD, etc.)

Techniques of ioTrace

- Work on top of block generic layer.
- Based on kernel blktrace API.
- Built with kernel tracepoints.



The API kernel provided

The statistics that ioTrace collects and manipulates:

```
struct blk_io_trace {
  __u32 magic;      /* MAGIC << 8 | version */
  __u32 sequence;   /* event number */
  __u64 time;   /* in nanoseconds */
  __u64 sector;     /* disk offset */
  __u32 bytes;      /* transfer length */
  __u32 action;     /* what happened */
  __u32 pid;    /* who did it */
  __u32 device;     /* device identifier (dev_t) */
  __u32 cpu;    /* on what cpu did it happen */
  __u16 error;      /* completion error */
  __u16 pdu_len;    /* length of data after this trace */
};
```

The stages of IO requests are represented by:

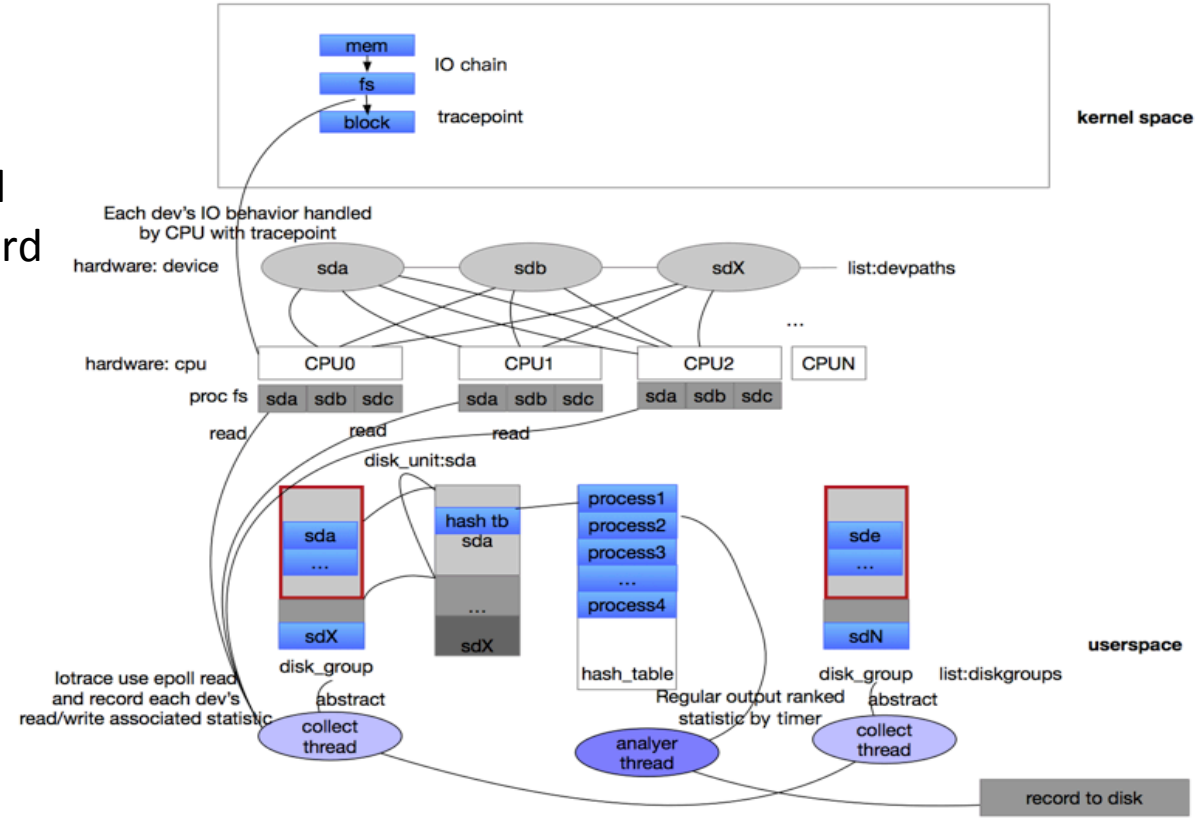
```
enum {
  BLK_TC_READ = 1 << 0,  /* reads */
  BLK_TC_WRITE  = 1 << 1,  /* writes */
  BLK_TC_FLUSH  = 1 << 2,  /* flush */
  BLK_TC_SYNC = 1 << 3,  /* sync */
  BLK_TC_QUEUE  = 1 << 4,  /* queueing/merging */
  BLK_TC_REQUEUE = 1 << 5,  /* requeueing */
  BLK_TC_ISSUE  = 1 << 6,  /* issue */
  BLK_TC_COMPLETE = 1 << 7,  /* completions */
  BLK_TC_FS  = 1 << 8,  /* fs requests */
  BLK_TC_PC  = 1 << 9,  /* pc requests */
  BLK_TC_NOTIFY  = 1 << 10,  /* special message */
  BLK_TC_AHEAD   = 1 << 11,  /* readahead */
  BLK_TC_META = 1 << 12,  /* metadata */
  BLK_TC_DISCARD  = 1 << 13,  /* discard requests */
  BLK_TC_DRV_DATA = 1 << 14,  /* binary driver data */
  BLK_TC_FUA = 1 << 15,  /* fua requests */

  BLK_TC_END  = 1 << 15,  /* we've run out of bits! */
};
```

The design of iotrace

Key objects and components:

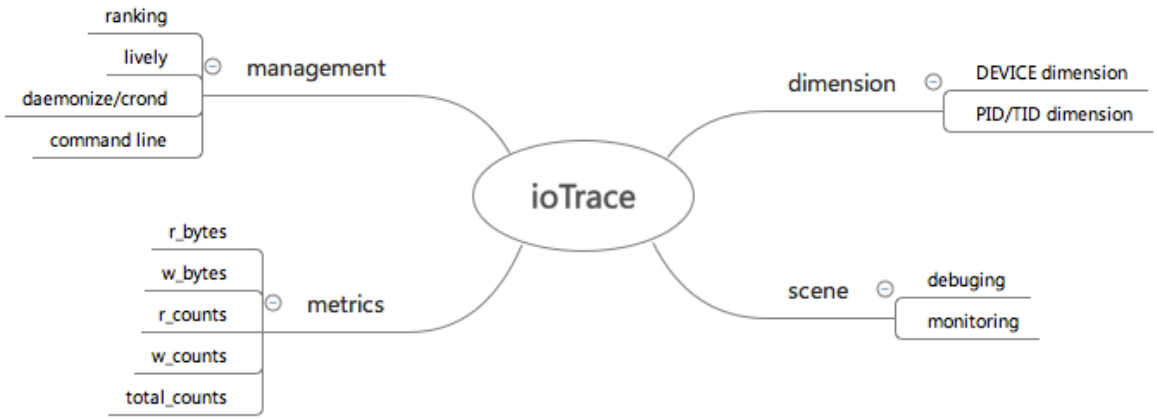
1. CPU List
2. Disk group
3. Epoll
4. Collect thread
5. Analyzer thread
6. Hash table record
7. Ranking logic



iotrace framework introduction

Functions of ioTrace

- Support TID, PID and DEVICE dimentions.
- Collect read_iops, write_iops, read_bytes, write_bytes, total_counts.
- Support prompt output to console and lagged json output to remote database.
- Support deamonizing and crond'ing mode with systemd.
- Support specifying target DEVICE name for monitoring.



Usage

Support multiple arguments: target device, prompt output mode, daemonizition or crond running mode, ranking output.

```
#iotrace
Usage: iotrace

[ -d <dev>            | --dev=<dev>                        ]
[ -m                  | --daemon                          ]
[ -c                  | --cron                            ]
[ -n <number>         | --top_candidates=<pid top max>]
[ -f <filename>       | --file=<configure file>            ]
[ -v <version>        | --version                          ]
[ -l <live>           | --live                             ]
[ -i <interval>       | --interval=<seconds>               ]
[ -p <thread>         | --thread=<count>                   ]

        -d Used to specify device
        -m Used to specify daemonize running or not
        -c Used to specify cron running or not
        -n Used to specify top candidates, defaults is 3
        -l Used to specify show data live or not
        -p Used to specify multiple thread max count
        -i Used to specify interval(second)
        -f Path to iotrace configure file, defaults to /etc/iotrace/iotrace.conf
        e.g:
                #./iotrace -d all -li1
                #./iotrace -d /dev/sda,/dev/sdc -li1
                #./iotrace -c
```

Data Accuracy

ioTrace

```
------------------------timestamp:2018-05-29 13:11:03--------device:sdc---------------------------------------
1 pid:112632  process:direct_io_        r_count:2890     w_count:0        r_bytes:2959360  w_bytes:0        t_count:2890
2 pid:0       process:                  r_count:0        w_count:0        r_bytes:0        w_bytes:0        t_count:0
3 pid:0       process:                  r_count:0        w_count:0        r_bytes:0        w_bytes:0        t_count:0
------------------------timestamp:2018-05-29 13:11:04--------device:sdc---------------------------------------
1 pid:112632  process:direct_io_        r_count:13542    w_count:0        r_bytes:13867008  w_bytes:0        t_count:13542
2 pid:0       process:                  r_count:0        w_count:0        r_bytes:0        w_bytes:0        t_count:0
3 pid:0       process:                  r_count:0        w_count:0        r_bytes:0        w_bytes:0        t_count:0
```

iostat

```
Device:         rrqm/s   wrqm/s     r/s     w/s    rkB/s    wkB/s avgrq-sz avgqu-sz   await r_await w_await  svctm  %util
sdc               0.00     0.00 2737.00    0.00  2737.00     0.00     2.00     0.23    0.08    0.08    0.00   0.08  22.90

Device:         rrqm/s   wrqm/s     r/s     w/s    rkB/s    wkB/s avgrq-sz avgqu-sz   await r_await w_await  svctm  %util
sdc               0.00     0.00 14052.00    0.00 14052.00     0.00     2.00     0.61    0.04    0.04    0.00   0.04  60.80

Device:         rrqm/s   wrqm/s     r/s     w/s    rkB/s    wkB/s avgrq-sz avgqu-sz   await r_await w_await  svctm  %util
sdc               0.00     0.00 3211.00    0.00  3211.00     0.00     2.00     0.26    0.08    0.08    0.00   0.08  26.30
```

Timestamp	Metric	ioTrace	iostat	Offset
20180529 13:11:03	r_bytes	2890KB	2737KB	+5.5%
20180529 13:11:04	r_bytes	13542KB	14052KB	-3.6%

Case

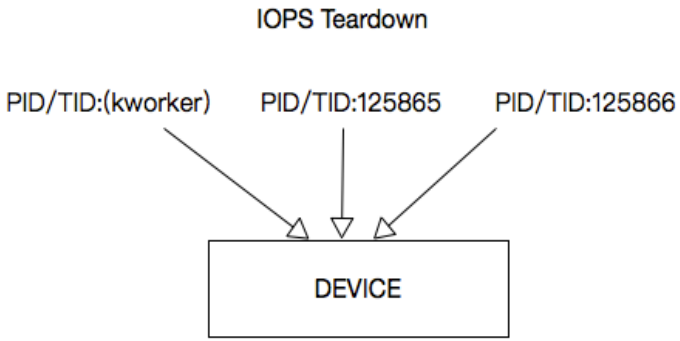
Output from ioTrace:

```
-------------------------timestamp:2018-06-20 19:48:05--------device:sdc-----------------------------------------
1 pid:125863  process:pangu_chun        r_count:63       w_count:26       r_bytes:15192064  w_bytes:12587008  t_count:89
2 pid:125864  process:pangu_chun        r_count:53       w_count:24       r_bytes:14225408  w_bytes:11636736  t_count:77
3 pid:125865  process:pangu_chun        r_count:47       w_count:16       r_bytes:13852672  w_bytes:6828032   t_count:63
-------------------------timestamp:2018-06-20 19:48:06--------device:sdc-----------------------------------------
1 pid:125865  process:pangu_chun        r_count:43       w_count:6        r_bytes:9924608   w_bytes:3129344   t_count:49
2 pid:125864  process:pangu_chun        r_count:46       w_count:1        r_bytes:11411456  w_bytes:520192    t_count:47
3 pid:125866  process:pangu_chun        r_count:41       w_count:2        r_bytes:10940416  w_bytes:1040384   t_count:43
-------------------------timestamp:2018-06-20 19:48:07--------device:sdc-----------------------------------------
1 pid:233962  process:kworker/u4        r_count:0        w_count:222      r_bytes:0         w_bytes:114864128 t_count:222
2 pid:125864  process:pangu_chun        r_count:51       w_count:0        r_bytes:11661312  w_bytes:0         t_count:51
3 pid:125865  process:pangu_chun        r_count:50       w_count:0        r_bytes:12619776  w_bytes:0         t_count:50
-------------------------timestamp:2018-06-20 19:48:08--------device:sdc-----------------------------------------
1 pid:233962  process:kworker/u4        r_count:2        w_count:166      r_bytes:8192      w_bytes:83873792  t_count:168
2 pid:125865  process:pangu_chun        r_count:40       w_count:0        r_bytes:10547200  w_bytes:0         t_count:40
3 pid:125866  process:pangu_chun        r_count:31       w_count:0        r_bytes:8232960   w_bytes:0         t_count:31
-------------------------timestamp:2018-06-20 19:48:09--------device:sdc-----------------------------------------
1 pid:233962  process:kworker/u4        r_count:3        w_count:288      r_bytes:12288     w_bytes:113545216 t_count:291
2 pid:125866  process:pangu_chun        r_count:15       w_count:0        r_bytes:4853760   w_bytes:0         t_count:15
3 pid:125864  process:pangu_chun        r_count:10       w_count:0        r_bytes:2174976   w_bytes:0         t_count:10
```

Output from SAR: disk util 100%

```
Time                 ----------------------------------------------------------sdc-----------------------------------------------------------
Time                rrqms   wrqms   %rrqm   %wrqm      rs      ws   rsecs   wsecs  rqsize  rarqsz  warqsz  qusize   await  rawait  wawait   svctm    util
20/06/18-19:47:52    0.00   59.00    0.00   29.80   71.00  139.00   46.5K  135.7K  444.25  335.66  499.71   69.00  458.16   21.37  681.27    4.70   98.70
20/06/18-19:47:53    0.00    0.00    0.00    0.00   51.00    4.00   34.0K    1.5K  330.98  341.57  196.00    0.00   13.84   11.29   46.25    5.85   32.20
20/06/18-19:47:54    0.00    0.00    0.00    0.00  125.00    5.00  106.4K    5.0K  438.74  435.87  510.40    1.00   14.91   10.30  130.20    4.38   57.00
20/06/18-19:47:55    0.00    0.00    0.00    0.00   71.00   12.00   65.7K   11.9K  479.18  474.08  509.33    0.00   10.45    9.93   13.50    5.72   47.50
20/06/18-19:47:56    0.00    0.00    0.00    0.00  102.00   44.00   75.7K   42.0K  412.74  379.92  488.82    3.00   20.79   11.64   42.00    4.61   67.30
20/06/18-19:47:57    0.00    0.00    0.00    0.00  130.00   50.00  107.5K   47.0K  439.64  423.57  481.44   82.00  114.32   28.07  338.56    5.57  100.00
20/06/18-19:47:58    0.00   52.00    0.00   83.87  162.00   10.00  119.5K    8.4K  380.86  377.75  431.20   76.00  109.96   32.35    1.3K    5.83  100.00
20/06/18-19:47:59    0.00    0.00    0.00    0.00  192.00   76.00  116.1K   74.0K  363.07  309.54  498.32   65.00  690.67   20.00    2.3K    3.74  100.00
20/06/18-19:48:00    0.00    0.00    0.00    0.00  198.00   10.00   99.6K    8.0K  264.87  257.64  408.00    1.00    9.60    9.22   17.10    2.99   62.20
20/06/18-19:48:01    0.00    0.00    0.00    0.00  122.00    6.00   85.2K    5.0K  360.62  357.44  425.33    2.00   16.28   16.06   20.83    5.32   68.10
20/06/18-19:48:02    0.00    0.00    0.00    0.00  108.00   14.00   82.8K   13.1K  402.43  392.41  479.71    2.00   23.89   16.33   82.21    5.71   69.70
20/06/18-19:48:03    0.00   43.00    0.00   50.59  131.00   42.00   83.3K   35.4K  351.21  325.40  431.71   16.00   95.82   18.62  336.62    5.32   92.00
20/06/18-19:48:04    0.00    0.00    0.00    0.00  216.00    7.00  109.0K    7.0K  266.15  258.28  509.14    2.00   12.30   11.57   35.00    3.71   82.80
20/06/18-19:48:05    0.00   44.00    0.00   33.85  183.00   86.00   91.0K   76.6K  318.87  254.47  455.91   38.00  141.40   28.66  381.30    3.73  100.00
20/06/18-19:48:06    0.00    1.00    0.00    6.67  180.00   14.00   83.6K   12.9K  254.66  237.71  472.57    3.00   15.13   14.04   29.14    3.32   64.50
20/06/18-19:48:07    1.00    0.00    0.63    0.00  157.00   93.00   72.4K   90.1K  332.75  236.10  495.91  144.00  129.74   23.17  309.66    4.01  100.00
20/06/18-19:48:08    0.00    1.00    0.00    0.55  124.00  180.00   61.7K  179.9K  406.83  254.61  511.69  141.00  678.27   18.19    1.1K    3.30  100.00
20/06/18-19:48:09    0.00    0.00    0.00    0.00   36.00  334.00   19.3K  261.5K  388.55  274.33  400.86  140.00  430.06   39.67  472.14    2.71  100.00
20/06/18-19:48:10    0.00   93.00    0.00   47.21   49.00  104.00   31.2K   89.9K  485.33  326.40  442.62   19.00  319.97   20.73  460.96    4.10   62.80
```

Consequence: Kworker is the obstacle



Case

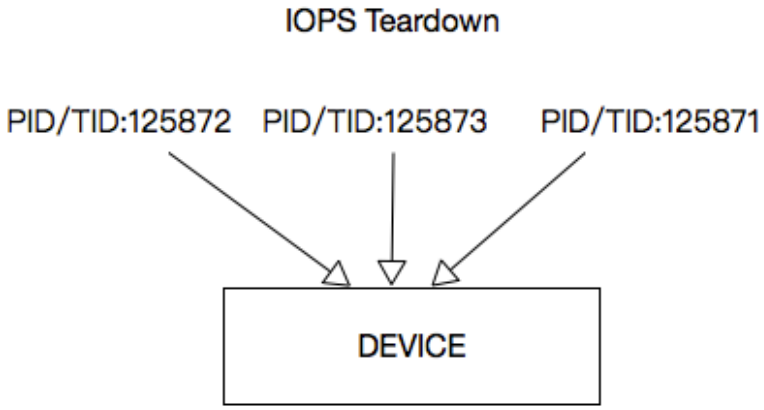
Output from ioTrace:

```
Time                 ---------------------------------------------------------sde----------------------------------------------------------
Time                rrqms   wrqms   %rrqm   %wrqm      rs      ws   rsecs   wsecs  rqsize  rarqsz  warqsz  qusize   await  rawait  wawait   svctm    util
20/06/18-20:35:41    0.00   22.00    0.00   22.00   29.00   78.00   12.3K   68.8K  388.07  217.79  451.38    4.00   44.61   17.52   54.68    4.08   43.70
20/06/18-20:35:42    0.00    0.00    0.00    0.00   38.00   19.00   23.2K   17.0K  361.40  312.84  458.53    0.00    5.44    7.45    1.42    4.56   26.00
^[[B20/06/18-20:35:43    0.00    0.00    0.00    0.00   66.00   69.00   28.7K   61.1K  340.39  222.55  453.10    3.00   24.44   11.24   37.06    4.13   55.80
20/06/18-20:35:44    0.00    0.00    0.00    0.00   37.00   18.00   22.1K   16.0K  354.40  305.19  455.56    1.00   21.76   23.11   19.00    8.25   45.40
20/06/18-20:35:45    0.00    0.00    0.00    0.00   34.00   93.00   22.5K   83.1K  425.51  338.35  457.38    1.00   14.32   18.65   12.74    4.34   55.10
20/06/18-20:35:46    0.00   75.00    0.00   29.88   28.00  176.00   21.4K   52.6K  185.82  391.86  153.05   41.00  116.65   25.68  131.12    2.88   58.70
20/06/18-20:35:47    0.00    0.00    0.00    0.00   45.00  138.00   27.9K  131.2K  445.07  317.42  486.70   45.00  348.72   31.40  452.20    4.85   88.80
20/06/18-20:35:48    0.00    0.00    0.00    0.00   42.00    7.00   10.4K    6.0K  171.10  126.38  439.43    0.00    9.94   11.31    1.71    4.04   19.80
20/06/18-20:35:49    0.00    0.00    0.00    0.00   46.00   37.00   22.1K   32.9K  339.33  246.43  454.81    1.00   16.89   13.09   21.62    5.08   42.20
20/06/18-20:35:50    0.00    0.00    0.00    0.00   32.00    1.00   11.0K    1.0K  185.82  175.62  512.00    0.00   10.33   10.62    1.00    5.00   16.50
20/06/18-20:35:51   12.00   41.00    8.00   87.23  138.00    6.00   45.8K    3.0K  173.67  169.97  258.67    3.00   20.42   19.86   33.33    5.44   78.30
20/06/18-20:35:52    0.00    0.00    0.00    0.00  268.00    0.00   57.8K    0.00  110.34  110.34    0.00    9.00   21.67   21.67    0.00    3.74  100.00
20/06/18-20:35:53    0.00    0.00    0.00    0.00  336.00   15.00   57.1K   10.1K   98.04   86.98  345.87    9.00   39.45   12.32  647.20    2.75   96.50
20/06/18-20:35:54    0.00    0.00    0.00    0.00   51.00    0.00   31.8K    0.00  318.82  318.82    0.00    0.00   10.51   10.51    0.00    6.49   33.10
```

Output from SAR:

```
-------------------------timestamp:2018-06-20 20:35:51--------device:sde-----------------------------------------
1 pid:125872  process:pangu_chun         r_count:41        w_count:0         r_bytes:5701632   w_bytes:0         t_count:41
2 pid:125873  process:pangu_chun         r_count:27        w_count:3         r_bytes:4538368   w_bytes:888832    t_count:30
3 pid:125871  process:pangu_chun         r_count:27        w_count:0         r_bytes:8085504   w_bytes:0         t_count:27
-------------------------timestamp:2018-06-20 20:35:52--------device:sde-----------------------------------------
1 pid:125872  process:pangu_chun         r_count:133       w_count:0         r_bytes:3473408   w_bytes:0         t_count:133
2 pid:125874  process:pangu_chun         r_count:63        w_count:0         r_bytes:14979072  w_bytes:0         t_count:63
3 pid:125871  process:pangu_chun         r_count:46        w_count:0         r_bytes:10317824  w_bytes:0         t_count:46
-------------------------timestamp:2018-06-20 20:35:53--------device:sde-----------------------------------------
1 pid:125872  process:pangu_chun         r_count:121       w_count:2         r_bytes:13074432  w_bytes:1048576   t_count:123
2 pid:125873  process:pangu_chun         r_count:112       w_count:5         r_bytes:5124096   w_bytes:2101248   t_count:117
3 pid:125871  process:pangu_chun         r_count:71        w_count:4         r_bytes:4153344   w_bytes:2097152   t_count:75
```

Consequence: PID 125872 is suspecious




LINUXCON
containercon
CLOUDOPEN
CHINA 中国
THINK OPEN
开放性思维
Thanks &
Questions
LF ASIA, LLC


LINUXCON
containercon
CLOUDOPEN
CHINA 中国
THINK OPEN
开放性思维
LF ASIA, LLC