eBPF在MySQL性能分析的应用

洪 斌

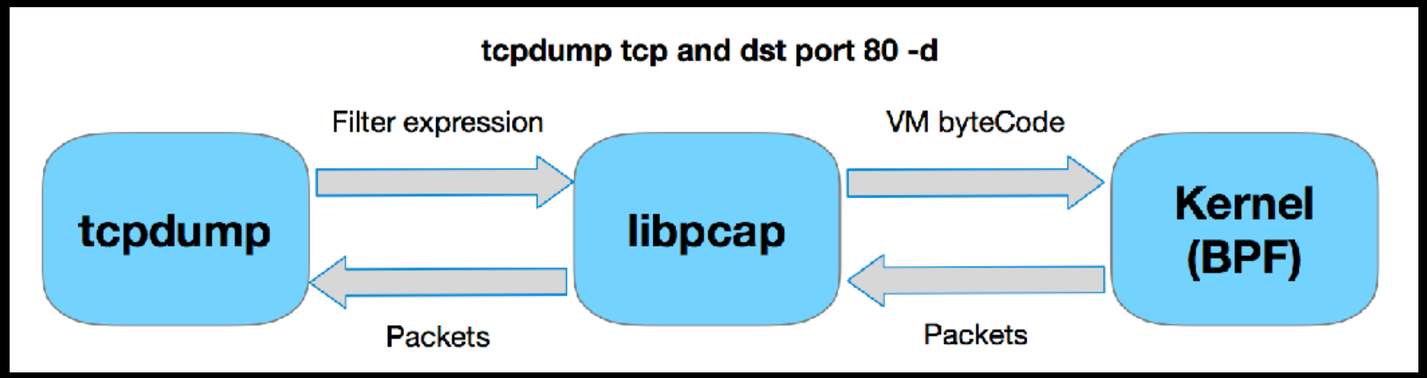
TECHNOLOGY ACTION 爱可生

目录

- eBPF 应用示例
- 观测工具的介绍
- eBPF 脚本/限制

BPF是什么

- BPF = Berkeley Packet Filter
- The Berkeley Packet Filter (BPF) provides a raw interface to data link layers, permitting raw link-layer packets to be sent and received.



- Since version 3.18, the Linux kernel includes an extended BPF virtual machine, termed extended BPF (eBPF). It can be used for non-networking purposes

http://www.tcpdump.org/papers/bpf-usenix93.pdf

Query延迟分布

```
//Only select

root@R820-08:/usr/share/bcc/tools# ./dbstat -p `pidof mysqld` -u -- mysql
Tracing database queries for pids 4754 slower than 0 ms...
^C[11:20:53]
     query latency (us)  : count     distribution
         0 -> 1          : 0        |                                        |
         2 -> 3          : 0        |                                        |
         4 -> 7          : 0        |                                        |
         8 -> 15         : 0        |                                        |
        16 -> 31         : 0        |                                        |
        32 -> 63         : 0        |                                        |
        64 -> 127        : 400308   |****************************************|
       128 -> 255        : 148021   |**************                          |
       256 -> 511        : 261      |                                        |
       512 -> 1023       : 3        |                                        |
      1024 -> 2047       : 0        |                                        |
      2048 -> 4095       : 1        |                                        |
      4096 -> 8191       : 3        |                                        |
      8192 -> 16383      : 9        |                                        |
```

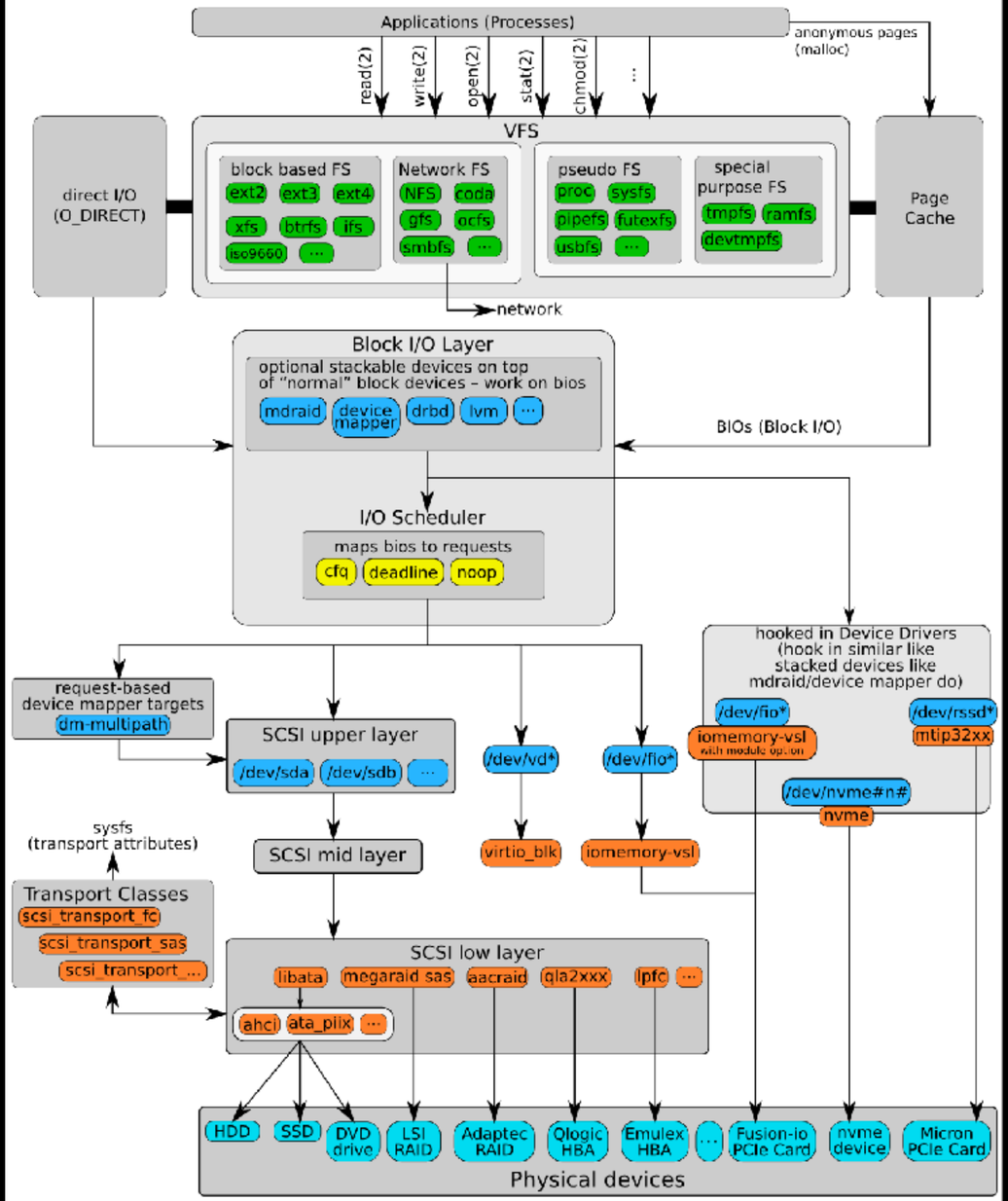
```
// Select and update
root@R820-08:/usr/share/bcc/tools# ./dbstat -p `pidof mysqld` -u -- mysql
Tracing database queries for pids 4754 slower than 0 ms...
^C[11:20:33]
     query latency (us)  : count     distribution
         0 -> 1          : 0        |                                        |
         2 -> 3          : 0        |                                        |
         4 -> 7          : 0        |                                        |
         8 -> 15         : 0        |                                        |
        16 -> 31         : 0        |                                        |
        32 -> 63         : 0        |                                        |
        64 -> 127        : 9198     |**************                          |
       128 -> 255        : 25826    |****************************************|
       256 -> 511        : 17629    |***************************             |
       512 -> 1023       : 14568    |**********************                  |
      1024 -> 2047       : 12533    |*******************                     |
      2048 -> 4095       : 9840     |***************                         |
      4096 -> 8191       : 4031     |******                                  |
      8192 -> 16383      : 463      |                                        |
     16384 -> 32767      : 33       |                                        |
     32768 -> 65535      : 20       |                                        |
     65536 -> 131071     : 20       |                                        |
```

慢Query抓取

```
// Select and update
root@R820-08:/usr/share/bcc/tools# ./dbslower -p `pidof mysqld` -m 5 -- mysql
Tracing database queries for pids 4754 slower than 5 ms...
TIME(s)        PID          MS QUERY
0.956044       4754      5.358 UPDATE sbtest1 SET k=k+1 WHERE id=514
0.956199       4754      5.837 UPDATE sbtest1 SET k=k+1 WHERE id=505
0.956876       4754      5.257 UPDATE sbtest1 SET k=k+1 WHERE id=503
0.955977       4754      6.656 UPDATE sbtest1 SET k=k+1 WHERE id=503
0.956287       4754      6.801 UPDATE sbtest1 SET k=k+1 WHERE id=503
0.955870       4754      7.554 UPDATE sbtest1 SET k=k+1 WHERE id=498
0.956329       4754      7.121 UPDATE sbtest1 SET k=k+1 WHERE id=497
...
```

The Linux I/O Stack Diagram

version 0.1, 2012-03-06
outlines the Linux I/O stack as of Kernel version 3.3



VFS 延迟分析

```
// Select and update
root@R730-117:/usr/share/bcc/tools# ./ext4dist 2 1
Tracing ext4 operation latency... Hit Ctrl-C to end.

21:39:52:

operation = read
             usecs                    : count     distribution
                 0 -> 1               : 0        |                    |
                 2 -> 3               : 3        |                    |
                 4 -> 7               : 19596    |***********         |
                 8 -> 15              : 32887    |********************|
                16 -> 31              : 2649     |*                   |
                32 -> 63              : 303      |                    |
                64 -> 127             : 48       |                    |
               128 -> 255             : 15       |                    |
               256 -> 511             : 3        |                    |
…
operation = write
             usecs                    : count     distribution
                 0 -> 1               : 0        |                    |
                 2 -> 3               : 2        |                    |
                 4 -> 7               : 507      |                    |
                 8 -> 15              : 22123    |********************|
                16 -> 31              : 10444    |*********           |
                32 -> 63              : 2073     |*                   |
                64 -> 127             : 590      |                    |
               128 -> 255             : 174      |                    |
               256 -> 511             : 240      |                    |
```

```
…
operation = fsync
             usecs                    : count     distribution
                 0 -> 1               : 166      |                    |
                 2 -> 3               : 291      |                    |
                 4 -> 7               : 446      |*                   |
                 8 -> 15              : 22       |                    |
                16 -> 31              : 3        |                    |
                32 -> 63              : 1        |                    |
                64 -> 127             : 2847     |*******             |
               128 -> 255             : 7164     |********************|
               256 -> 511             : 4292     |***********         |
               512 -> 1023            : 882      |**                  |
```

Ext4 延迟分析

```
//Insert data

root@R820-08:/usr/share/bcc/tools# ./ext4slower 1
Tracing ext4 operations slower than 1 ms
TIME     COMM           PID    T BYTES   OFF_KB   LAT(ms) FILENAME
21:59:40 mysqld         4754   S 0       0           3.56 ib_logfile1
21:59:40 mysqld         4754   S 0       0           8.42 sbtest1.ibd
21:59:41 mysqld         4754   S 0       0           3.83 ib_logfile1
21:59:41 mysqld         4754   S 0       0           8.35 sbtest1.ibd
21:59:42 mysqld         4754   S 0       0           8.50 sbtest1.ibd
21:59:42 mysqld         4754   S 0       0           3.53 ib_logfile1
21:59:42 mysqld         4754   S 0       0           8.34 sbtest1.ibd
21:59:43 mysqld         4754   S 0       0           2.69 ib_logfile1
21:59:43 mysqld         4754   S 0       0           8.41 sbtest1.ibd
21:59:44 mysqld         4754   S 0       0           8.37 sbtest1.ibd
21:59:44 mysqld         4754   S 0       0           4.13 ib_logfile1
21:59:44 mysqld         4754   S 0       0           8.38 sbtest1.ibd
21:59:45 mysqld         4754   S 0       0           8.52 sbtest1.ibd
```

```
root@R820-08:/usr/share/bcc/tools# ./ext4slower 10
Tracing ext4 operations slower than 10 ms
TIME     COMM           PID    T BYTES      OFF_KB   LAT(ms) FILENAME
22:03:14 dd             42639  W 1073741824 0         873.20 test1.img
22:03:15 mysqld         4754   W 1048576    1024       16.48 ibdata1
22:03:15 mysqld         4754   W 507904     2048       13.98 ibdata1
22:03:15 mysqld         4754   W 1048576    1302528    15.10 sbtest1.ibd
22:03:15 mysqld         4754   S 0          0         110.94 ibdata1
22:03:16 mysqld         4754   W 1048576    1306624    22.35 sbtest1.ibd
```

块设备延迟分析

```
//Select and update

root@R730-117:/usr/share/bcc/tools# ./biolatency -D 2
Tracing block device I/O... Hit Ctrl-C to end.


disk = 'sdb'
     usecs               : count     distribution
         0 -> 1          : 0        |                                        |
         2 -> 3          : 0        |                                        |
         4 -> 7          : 0        |                                        |
         8 -> 15         : 0        |                                        |
        16 -> 31         : 0        |                                        |
        32 -> 63         : 4694     |**********                              |
        64 -> 127        : 3399     |*******                                 |
       128 -> 255        : 2211     |*****                                   |
       256 -> 511        : 2250     |*****                                   |
       512 -> 1023       : 642      |*                                       |
      1024 -> 2047       : 0        |                                        |
      2048 -> 4095       : 0        |                                        |
```

```
root@R730-117:/usr/share/bcc/tools# ./biolatency -D 2
Tracing block device I/O... Hit Ctrl-C to end.


disk = 'sdb'
     usecs               : count     distribution
         0 -> 1          : 0        |                                        |
         2 -> 3          : 0        |                                        |
         4 -> 7          : 0        |                                        |
         8 -> 15         : 0        |                                        |
        16 -> 31         : 0        |                                        |
        32 -> 63         : 0        |                                        |
        64 -> 127        : 0        |                                        |
       128 -> 255        : 0        |                                        |
       256 -> 511        : 2        |**************************              |
       512 -> 1023       : 0        |                                        |
      1024 -> 2047       : 0        |                                        |
      2048 -> 4095       : 3        |****************************************|
```

MySQL文件IO压力分析

```
root@R820-08:/usr/share/bcc/tools# ./filetop -p `pidof mysqld` -C 5
Tracing... Output every 5 secs. Hit Ctrl-C to end

22:26:30 loadavg: 7.50 5.28 4.87 18/1925 44235

TID    COMM             READS  WRITES R_Kb    W_Kb    T FILE
39956  mysqld           0      115    0       462     R ib_logfile1
40075  mysqld           0      107    0       424     R ib_logfile1
39900  mysqld           0      1220   0       137     R R820-08.log
38046  mysqld           0      1263   0       142     R R820-08.log
39085  mysqld           0      101    0       332     R ib_logfile1
38957  mysqld           0      114    0       425     R ib_logfile1
39959  mysqld           0      1      0       2       R ibmPAQIO
4780   mysqld           0      4      0       28      R ib_logfile1
40266  mysqld           0      107    0       361     R ib_logfile1
39984  mysqld           0      111    0       414     R ib_logfile1
39991  mysqld           0      1211   0       136     R R820-08.log
37224  mysqld           0      104    0       449     R ib_logfile1
40259  mysqld           0      109    0       340     R ib_logfile1
39958  mysqld           0      107    0       342     R ib_logfile1
39969  mysqld           0      1214   0       137     R R820-08.log
39966  mysqld           0      1275   0       144     R R820-08.log
39937  mysqld           0      1227   0       138     R R820-08.log
```

临时表文件生命周期观测

```
root@R820-08:/usr/share/bcc/tools# ./filelife
TIME     PID    COMM                    AGE(s)  FILE
22:17:01 43687  cron                    0.00    tmpfgHF5vY
22:22:21 39170  mysqld                  5.30    #sql1292_59a1f_0.frm
```

短连接分析

```
root@R820-08:/usr/share/bcc/tools# ./tcplife
PID   COMM       LADDR           LPORT RADDR           RPORT TX_KB RX_KB MS
44245 sysbench   127.0.0.1       35038 127.0.0.1       3306     16   699 312.05
44245 sysbench   127.0.0.1       35036 127.0.0.1       3306     17   736 312.20
44245 sysbench   127.0.0.1       35034 127.0.0.1       3306     15   662 312.41
44245 sysbench   127.0.0.1       35032 127.0.0.1       3306     14   638 312.45
44245 sysbench   127.0.0.1       35026 127.0.0.1       3306     14   626 313.17
44245 sysbench   127.0.0.1       35028 127.0.0.1       3306     12   552 313.18
44245 sysbench   127.0.0.1       35022 127.0.0.1       3306     17   736 313.66
44245 sysbench   127.0.0.1       35018 127.0.0.1       3306     13   589 313.86
44245 sysbench   127.0.0.1       35016 127.0.0.1       3306     13   589 314.00
44245 sysbench   127.0.0.1       35014 127.0.0.1       3306     14   626 314.11
44245 sysbench   127.0.0.1       35012 127.0.0.1       3306     17   761 314.15
44245 sysbench   127.0.0.1       35010 127.0.0.1       3306     17   736 314.60
44245 sysbench   127.0.0.1       35008 127.0.0.1       3306     15   663 314.66
44245 sysbench   127.0.0.1       35004 127.0.0.1       3306     16   699 314.74
44245 sysbench   127.0.0.1       35002 127.0.0.1       3306     15   663 315.05
44245 sysbench   127.0.0.1       35000 127.0.0.1       3306     15   699 315.09
```

```
13:08:05.737768 ppp0 > slip139-92-26-177.ist.tr.ibm.net.1221 > dsl-usw-cust-110.inetarena.com.www: . 342:342(0) ack 1449 win 31856 <nop
,nop,timestamp 1247771 114849487> (DF)
13:08:07.467571 ppp0 < dsl-usw-cust-110.inetarena.com.www > slip139-92-26-177.ist.tr.ibm.net.1221: . 1449:2897(1448) ack 342 win 31856
<nop,nop,timestamp 114849637 1247771> (DF)
13:08:07.707634 ppp0 < dsl-usw-cust-110.inetarena.com.www > slip139-92-26-177.ist.tr.ibm.net.1221: . 2897:4345(1448) ack 342 win 31856
<nop,nop,timestamp 114849637 1247771> (DF)
13:08:07.707922 ppp0 > slip139-92-26-177.ist.tr.ibm.net.1221 > dsl-usw-cust-110.inetarena.com.www: . 342:342(0) ack 4345 win 31856 <nop
,nop,timestamp 1247968 114849637> (DF)
13:08:08.057841 ppp0 > slip139-92-26-177.ist.tr.ibm.net.1045 > ns.de.ibm.net.domain: 8928+ PTR? 110.107.102.209.in-addr.arpa. (46)
13:08:08.747598 ppp0 < dsl-usw-cust-110.inetarena.com.www > slip139-92-26-177.ist.tr.ibm.net.1221: P 4345:5793(1448) ack 342 win 31856
<nop,nop,timestamp 114849813 1247968> (DF)
13:08:08.847870 ppp0 < dsl-usw-cust-110.inetarena.com.www > slip139-92-26-177.ist.tr.ibm.net.1221: FP 5793:6297(504) ack 342 win 31856
<nop,nop,timestamp 114849813 1247968> (DF)
13:08:08.848063 ppp0 > slip139-92-26-177.ist.tr.ibm.net.1221 > dsl-usw-cust-110.inetarena.com.www: . 342:342(0) ack 6298 win 31856 <nop
,nop,timestamp 1248082 114849813> (DF)
13:08:08.907566 ppp0 < ns.de.ibm.net.domain > slip139-92-26-177.ist.tr.ibm.net.1045: 8928* 3/1/1 PTR dsl-usw-cust-110.inetarena.com., P
TR fingerless.or (199)
13:08:09.151742 ppp0 > slip139-92-26-177.ist.tr.ibm.net.1221 > dsl-usw-cust-110.inetarena.com.www: F 342:342(0) ack 6298 win 31856 <nop
,nop,timestamp 1248112 114849813> (DF)
13:08:10.137603 ppp0 < dsl-usw-cust-110.inetarena.com.www > slip139-92-26-177.ist.tr.ibm.net.1221: . 6298:6298(0) ack 343 win 31856 <no
p,nop,timestamp 114849967 1248112> (DF)
13:09:01.984210 ppp0 > slip139-92-26-177.ist.tr.ibm.net.1222 > dsl-usw-cust-110.inetarena.com.www: S 920197285:920197285(0) win 32120 <
mss 1460,sackOK,timestamp 1253395 0,nop,wscale 0> (DF)
13:09:03.097569 ppp0 < dsl-usw-cust-110.inetarena.com.www > slip139-92-26-177.ist.tr.ibm.net.1222: S 1222277738:1222277738(0) ack 92019
7286 win 32120 <mss 1460,sackOK,timestamp 114855252 1253395,nop,wscale 0> (DF)
13:09:03.098197 ppp0 > slip139-92-26-177.ist.tr.ibm.net.1222 > dsl-usw-cust-110.inetarena.com.www: . 1:1(0) ack 1 win 32120 <nop,nop,ti
mestamp 1253507 114855252> (DF)
13:09:03.102171 ppp0 > slip139-92-26-177.ist.tr.ibm.net.1222 > dsl-usw-cust-110.inetarena.com.www: P 1:322(321) ack 1 win 32120 <nop,no
p,timestamp 1253507 114855252> (DF)
13:09:04.147613 ppp0 < dsl-usw-cust-110.inetarena.com.www > slip139-92-26-177.ist.tr.ibm.net.1222: . 1:1(0) ack 322 win 31856 <nop,nop,
timestamp 114855369 1253507> (DF)
13:09:04.507608 ppp0 < dsl-usw-cust-110.inetarena.com.www > slip139-92-26-177.ist.tr.ibm.net.1222: . 1:1449(1448) ack 322 win 31856 <no
p,nop,timestamp 114855369 1253507> (DF)
13:09:04.507934 ppp0 > slip139-92-26-177.ist.tr.ibm.net.1222 > dsl-usw-cust-110.inetarena.com.www: . 322:322(0) ack 1449 win 31856 <nop
,nop,timestamp 1253648 114855369> (DF)
13:09:05.627604 ppp0 < dsl-usw-cust-110.inetarena.com.www > slip139-92-26-177.ist.tr.ibm.net.1222: . 1449:2897(1448) ack 322 win 31856
<nop,nop,timestamp 114855491 1253648> (DF)
13:09:05.857649 ppp0 < dsl-usw-cust-110.inetarena.com.www > slip139-92-26-177.ist.tr.ibm.net.1222: . 2897:4345(1448) ack 322 win 31856
<nop,nop,timestamp 114855491 1253648> (DF)
13:09:05.857918 ppp0 > slip139-92-26-177.ist.tr.ibm.net.1222 > dsl-usw-cust-110.inetarena.com.www: . 322:322(0) ack 4345 win 31856 <nop
,nop,timestamp 1253783 114855491> (DF)
13:09:06.907557 ppp0 < dsl-usw-cust-110.inetarena.com.www > slip139-92-26-177.ist.tr.ibm.net.1222: FP 4345:5792(1447) ack 322 win 31856
 <nop,nop,timestamp 114855627 1253783> (DF)
13:09:06.907887 ppp0 > slip139-92-26-177.ist.tr.ibm.net.1222 > dsl-usw-cust-110.inetarena.com.www: . 322:322(0) ack 5793 win 31856 <nop
,nop,timestamp 1253888 114855627> (DF)
13:09:07.401205 ppp0 > slip139-92-26-177.ist.tr.ibm.net.1222 > dsl-usw-cust-110.inetarena.com.www: F 322:322(0) ack 5793 win 31856 <nop
,nop,timestamp 1253937 114855627> (DF)
13:09:08.317623 ppp0 < dsl-usw-cust-110.inetarena.com.www > slip139-92-26-177.ist.tr.ibm.net.1222: . 5793:5793(0) ack 323 win 31856 <no
p,nop,timestamp 114855780 1253937> (DF)
```

流量分析

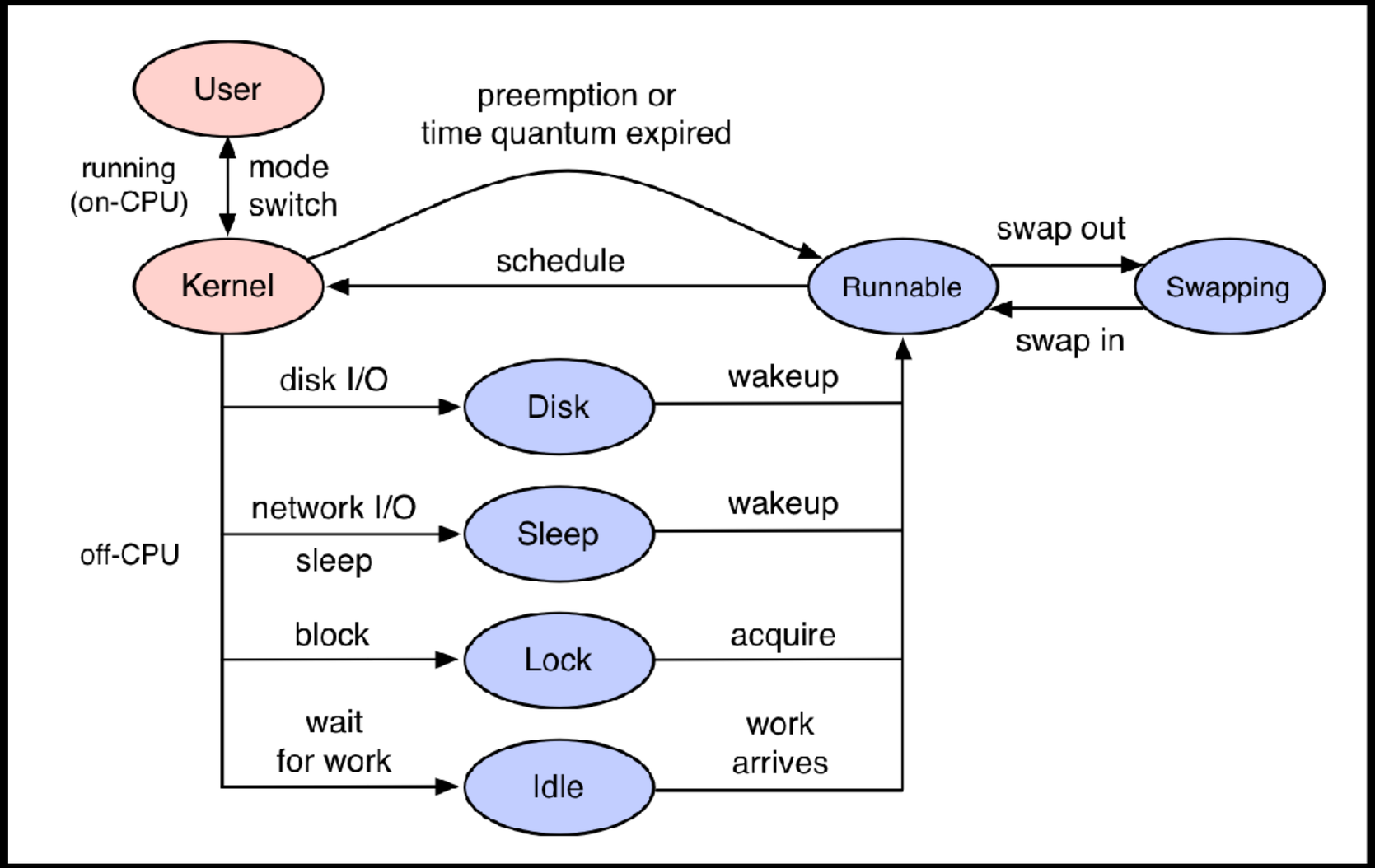
```
root@R820-08:/usr/share/bcc/tools# ./tcptop -C 5
Tracing... Output every 5 secs. Hit Ctrl-C to end
22:33:41 loadavg: 17.28 6.81 5.01 126/1933 44788

PID    COMM         LADDR                 RADDR                 RX_KB  TX_KB
44668  sysbench     127.0.0.1:35654       127.0.0.1:3306        16116    369
44669  sysbench     127.0.0.1:35650       127.0.0.1:3306        15957    365
44702  sysbench     127.0.0.1:35728       127.0.0.1:3306        15871    363
44758  sysbench     127.0.0.1:35838       127.0.0.1:3306        15834    362
44698  sysbench     127.0.0.1:35718       127.0.0.1:3306        15797    362
…


PID    COMM         LADDR6                          RADDR6
RX_KB  TX_KB
39188  mysqld       ::ffff:127.0.0.1:3306           ::ffff:127.0.0.1:35654
410  17810
39184  mysqld       ::ffff:127.0.0.1:3306           ::ffff:127.0.0.1:35682
405  17589
40037  mysqld       ::ffff:127.0.0.1:3306           ::ffff:127.0.0.1:35890
404  17552
39182  mysqld       ::ffff:127.0.0.1:3306           ::ffff:127.0.0.1:35670
404  17491
39180  mysqld       ::ffff:127.0.0.1:3306           ::ffff:127.0.0.1:35688
404  17540
...
```

跟踪函数调用

```
[root@hadoop03 tools]./trace -U -p 26720 '/usr/local/mysql/bin/mysqld:_Z24log_buffer_flush_to_diskb'
PID     TID     COMM            FUNC
26720   230362  mysqld          _Z24log_buffer_flush_to_diskb
        log_buffer_flush_to_disk(bool)+0x0 [mysqld]
        handler::ha_create(char const*, TABLE*, st_ha_create_information*)+0x74 [mysqld]
        ha_create_table(THD*, char const*, char const*, char const*, st_ha_create_information*, bool, bool)+0x241 [mysqld]
        rea_create_table(THD*, char const*, char const*, char const*, st_ha_create_information*, List<Create_field>&, unsigned int, st_key*,
handler*, bool)+0x1dd [mysqld]
        create_table_impl(THD*, char const*, char const*, char const*, char const*, st_ha_create_information*, Alter_info*, bool, unsigned
int, bool, bool*, st_key**, unsigned int*)+0x1498 [mysqld]
        mysql_create_table_no_lock(THD*, char const*, char const*, st_ha_create_information*, Alter_info*, unsigned int, bool*)+0x17c
[mysqld]
        mysql_create_table(THD*, TABLE_LIST*, st_ha_create_information*, Alter_info*)+0xf4 [mysqld]
        mysql_execute_command(THD*, bool)+0x1de4 [mysqld]
        mysql_parse(THD*, Parser_state*)+0x5fc [mysqld]
        dispatch_command(THD*, COM_DATA const*, enum_server_command)+0xca9 [mysqld]
        do_command(THD*)+0x4b2 [mysqld]
        handle_connection+0x1e0 [mysqld]
        pfs_spawn_thread+0x170 [mysqld]
        start_thread+0xc5 [libpthread-2.17.so]
```

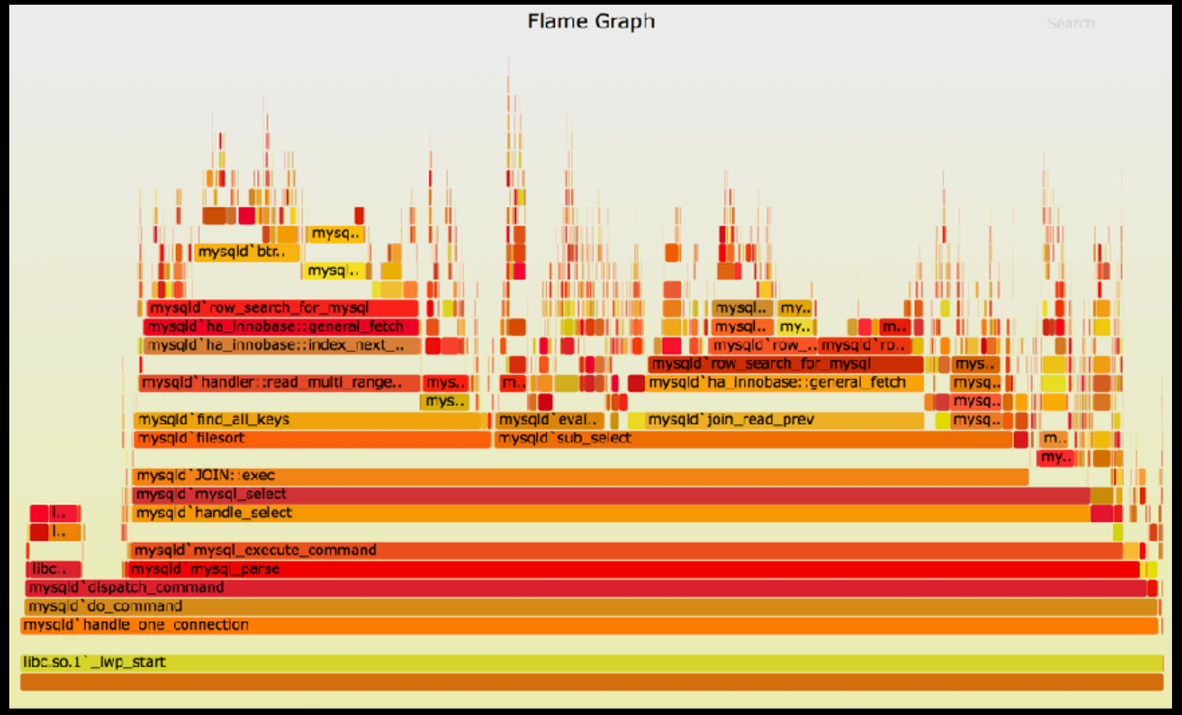
```
[root@hadoop03 ~]# objdump -t /usr/local/mysql/bin/mysqld|grep -i log_buffer_flush_to_disk
00000000019523a5 g     F .text    0000000000000057              _Z24log_buffer_flush_to_diskb
```

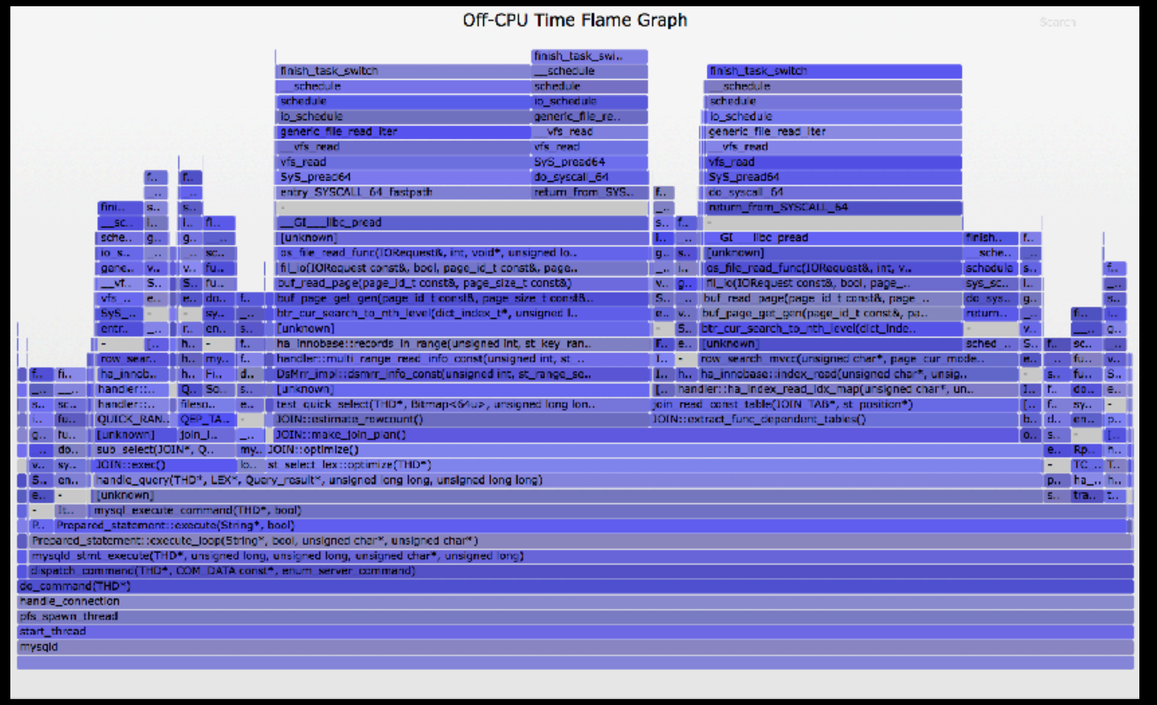
CPU消耗分析


User
preemption or
time quantum expired
running
(on-CPU)
mode
switch
swap out
schedule
Kernel
Runnable
Swapping
swap in
disk I/O
Disk
wakeup
network I/O
Sleep
wakeup
off-CPU
sleep
block
Lock
acquire
wait
for work
work
arrives
Idle

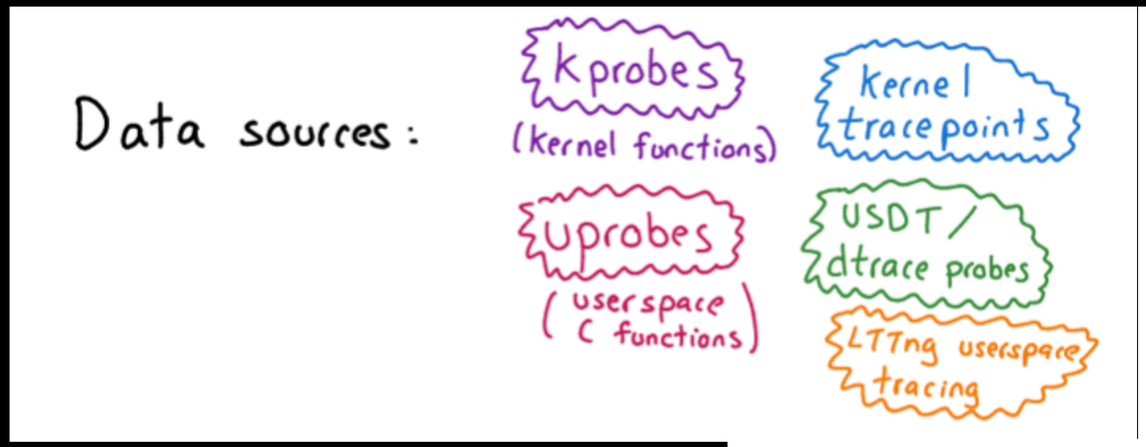
On/Off-CPU火焰图分析

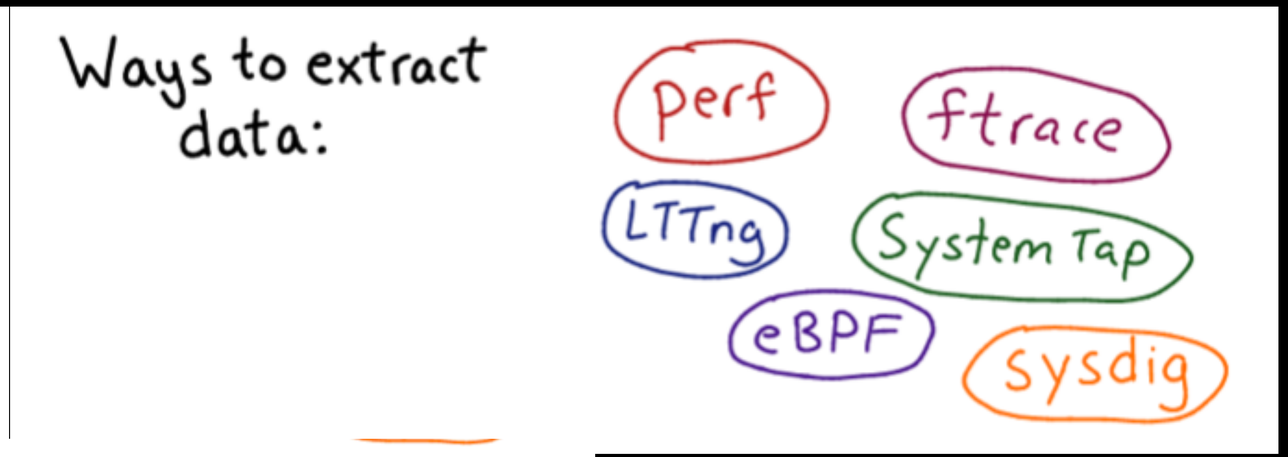
```
# /usr/share/bcc/tools/offcputime -df -p `pgrep -nx mysqld` 30 > out.stacks
[...copy out.stacks to your local system if desired...]
# git clone https://github.com/brendangregg/FlameGraph
# cd FlameGraph
# ./flamegraph.pl --color=io --title="Off-CPU Time Flame Graph" --countname=us < out.stacks > out.svg
```

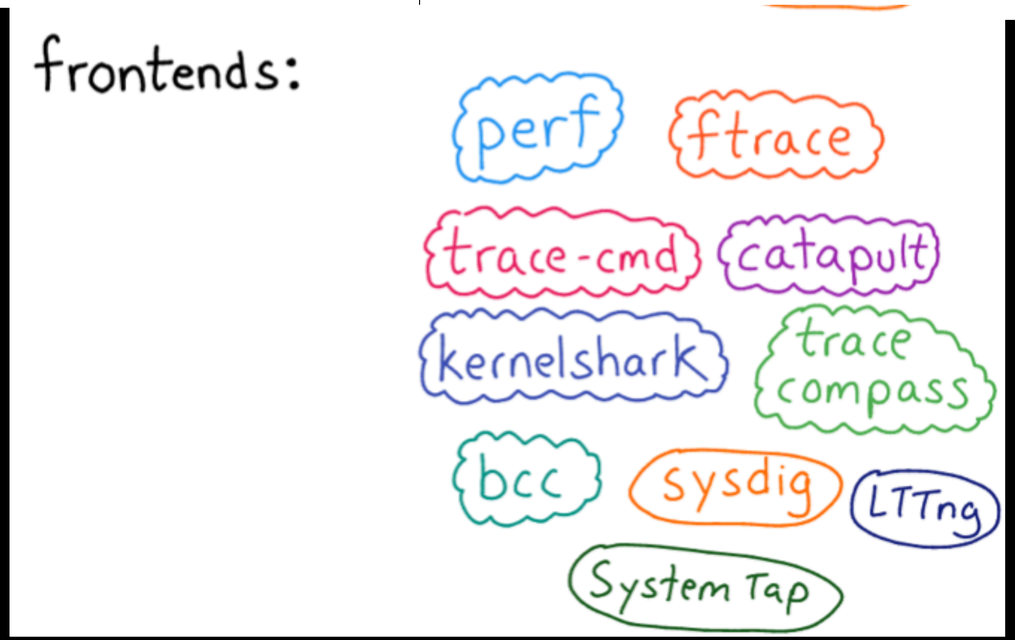




观测工具介绍


Data sources:
kprobes
(kernel functions)
kernel tracepoints
uprobes
(userspace C functions)
USDT / dtrace probes
LTTng userspace tracing

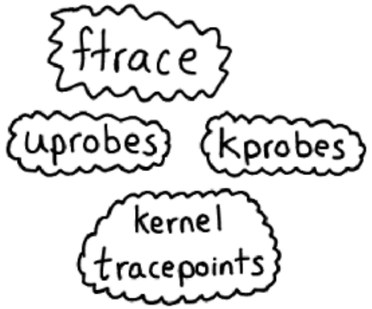

Ways to extract data:
perf
ftrace
LTTng
System Tap
eBPF
sysdig


frontends:
perf
ftrace
trace-cmd
catapult
kernelshark
trace compass
bcc
sysdig
LTTng
System Tap

eBPF vs Other

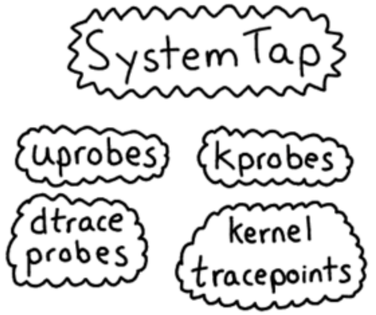
Ways to get (delicious delicious) tracing data

There are a bunch of ways to collect tracing data. These 3 are the ones that are built into the Linux kernel.

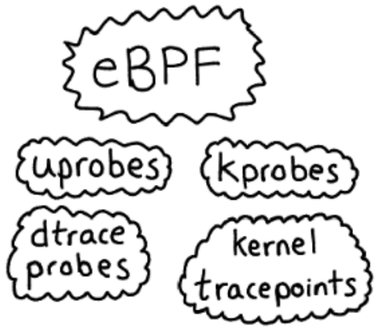


magical filesystem at /sys/kernel/debug/tracing. Super powerful, you interact with it by reading from / writing to files.

① call the perf_event_open syscall
② the kernel writes data to a ring buffer ("perf buffer")



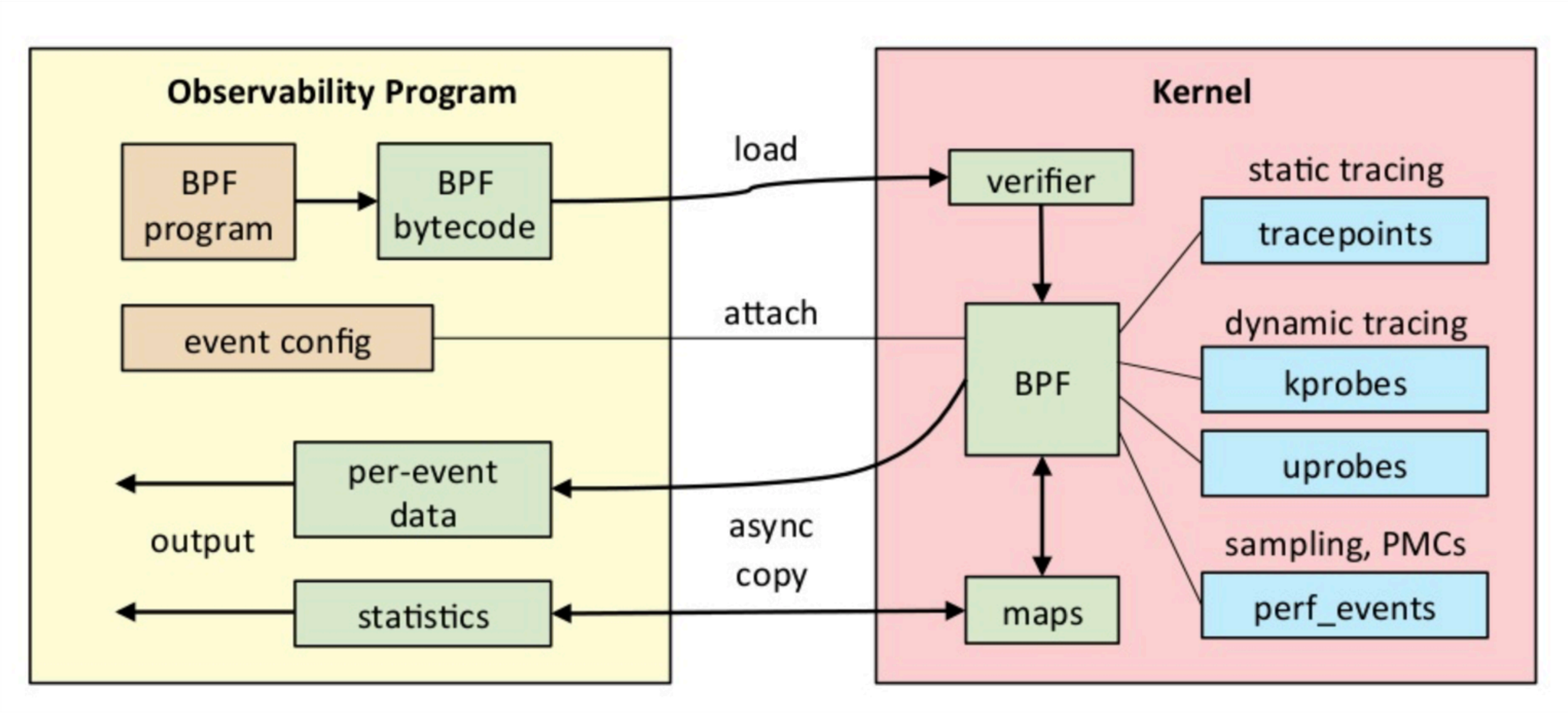
① Write some C code
② Compile it into a custom kernel module
③ Insert that module into the kernel



The newest and most powerful

① Write a small eBPF program
② Ask Linux to attach it to a kprobe/uprobe/tracepoint
③ The eBPF program sends data to userspace with ftrace/perf/BPF maps

eBPF内部


Observability Program
BPF program
BPF bytecode
event config
per-event data
output
statistics
load
attach
async copy
Kernel
verifier
BPF
maps
static tracing
tracepoints
dynamic tracing
kprobes
uprobes
sampling, PMCs
perf_events

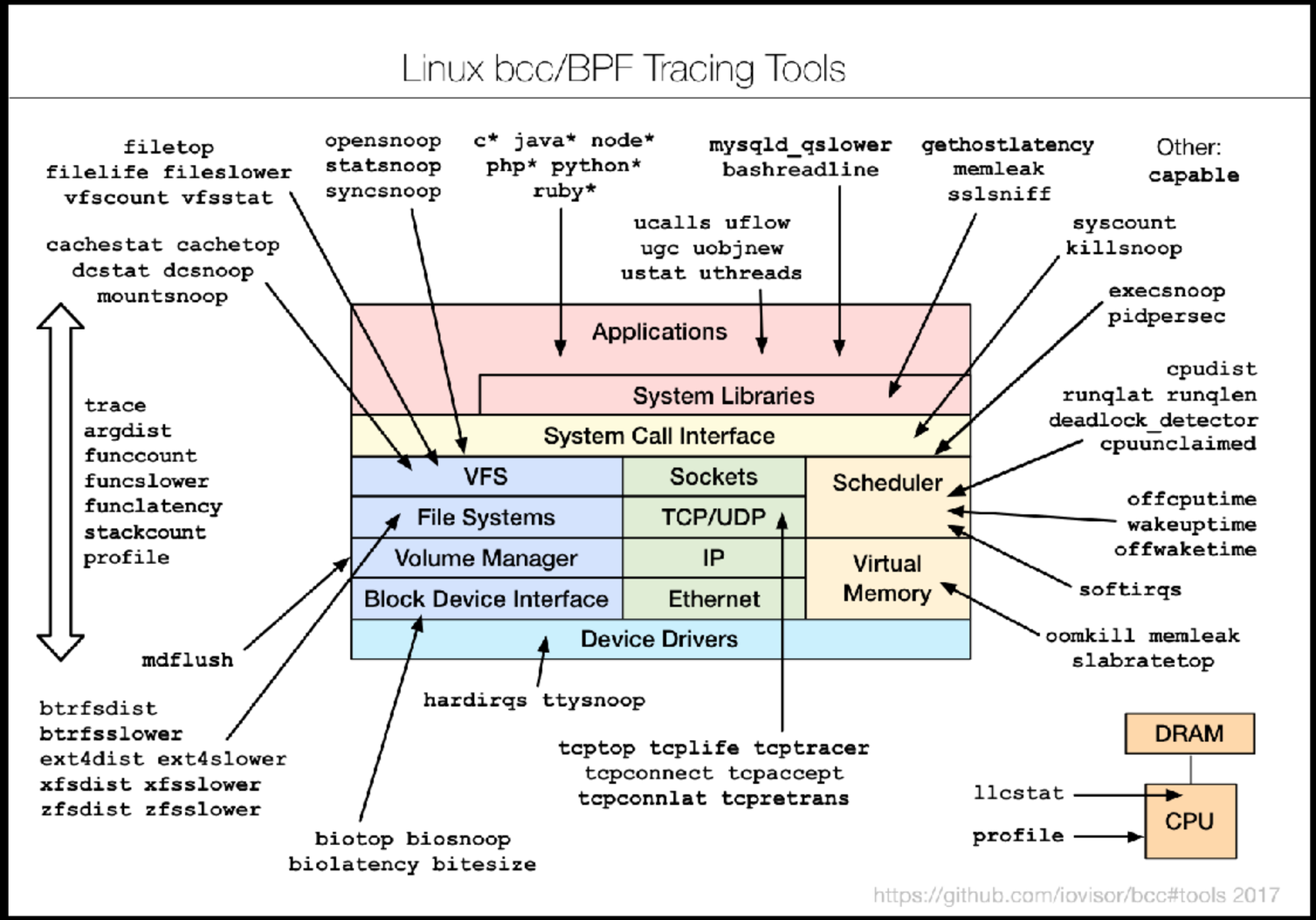
eBPF脚本

```
//一段C++代码，嵌入kprobe/uprobe
program = """
#include <uapi/linux/ptrace.h>

BPF_HASH(temp, u64, u64); //临时容器
BPF_HISTOGRAM(latency); //存放结果的容器

int probe_start(struct pt_regs *ctx) {
    u64 timestamp = bpf_ktime_get_ns(); //快速获取时间戳
    u64 pid = bpf_get_current_pid_tgid();
    temp.update(&pid, &timestamp);
    return 0;
}

int probe_end(struct pt_regs *ctx) {
    u64 *timestampp;
    u64 pid = bpf_get_current_pid_tgid();
    timestampp = temp.lookup(&pid);
    if (!timestampp)
        return 0;

    u64 delta = bpf_ktime_get_ns() - *timestampp; //获取时间差
    FILTER
    delta /= SCALE; //规范化时间差
    latency.increment(bpf_log2l(delta)); //存放结果
    temp.delete(&pid);
    return 0;
}
"""
```

```
...
// 将代码嵌入uprobe
usdts = map(lambda pid: USDT(pid=pid), args.pids)
for usdt in usdts:
    usdt.enable_probe("query__start", "probe_start")
    usdt.enable_probe("query__done", "probe_end")
bpf = BPF(text=program, usdt_contexts=usdts)
...
// 获取结果集
latencies = bpf["latency"]
...
// 打印结果集
latencies.print_log2_hist("query latency (%s)" %
                          ("us" if args.microseconds else "ms"))
...
```

eBPF 工具箱

1. execsnoop
2. opensnoop
3. ext4slower
4. biolatency
5. biosnoop
6. cachestat
7. tcpconnect
8. tcpaccept
9. tcpretrans
10. gethostlatency
11. runlat
12. profile



eBPF限制

- OS kernel 4.4+ (推荐 4.9+)
- MySQL 编译 -DENABLE_DTRACE=1 & 安装 systemtap-sdt-devel

引用

- https://lwn.net/Articles/603983/
- https://github.com/iovisor/bcc
- http://www.brendangregg.com/offcpuanalysis.html
- https://jvns.ca/blog/2017/07/05/linux-tracing-systems/#dtrace-probes
- http://www.brendangregg.com/ebpf.html

- 是某个慢还是所有都慢

Thanks

TECHNOLOGY
ACTION
爱可生