App Trouble Shooting On Linux

geekidea@gmail.com
2018 年 4 月

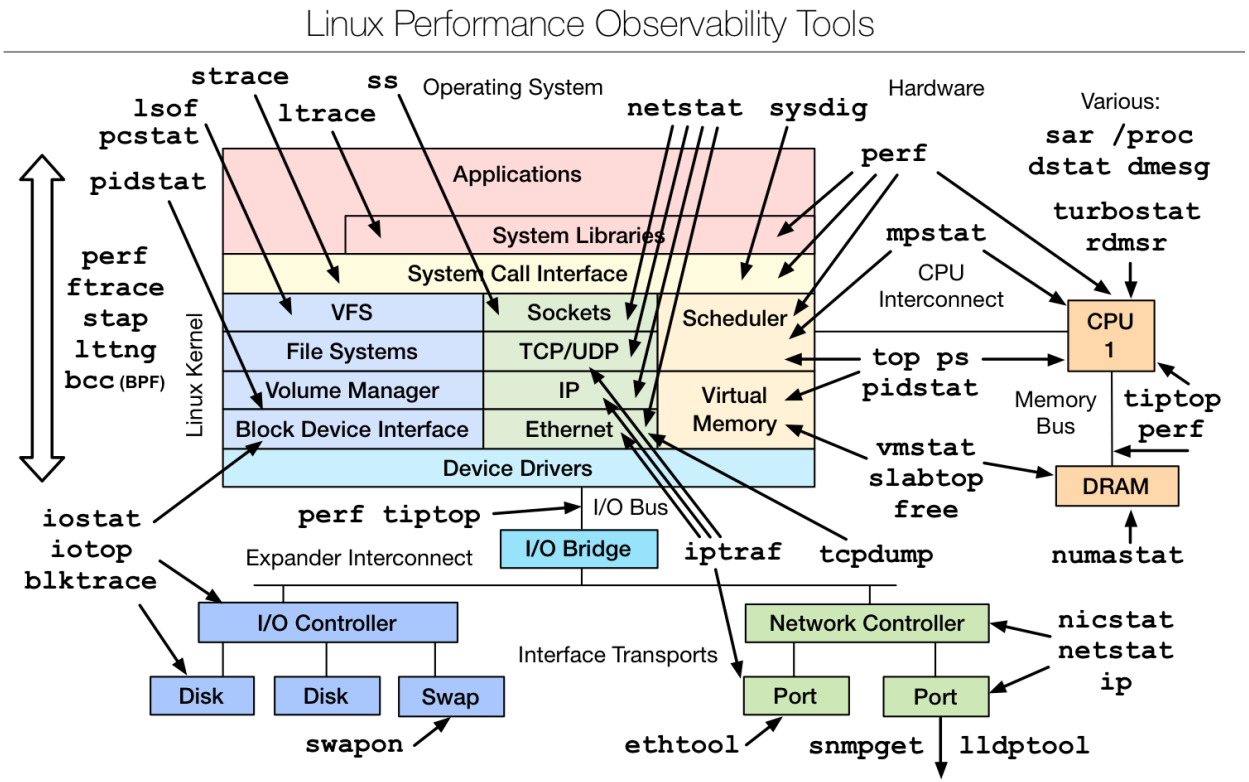
目录

- 常用方法技巧
- 常用命令集
- 问题分析举例
- Q&A

常用方法技巧

- 通过系统级别的方法先查看整体系统负载情况，再快速定位具体模块问题
- top / iotop / iftop等查看CPU负载，io负载，网卡负载总体情况
- ps / lsof查看（连接句柄）进程当前关联的句柄（进程）
- strace抓一段时间可疑进程的系统调用情况，分析strace log
- 可以查找：ERROR或其他关键字、对应文件或连接句柄的生命周期

常用命令集



http://www.brendangregg.com/linuxperf.html 2017

lsof

- 枚举进程打开的句柄（文件、sockets）
- 查看某一文件或连接（TCP/UDP端口）所关联的进程
- 查看flock文件锁&线程锁交叉导致的死锁(W)
- lsof
- lsof -nNPp <pid>
- lsof -nPi@192.168.16.31:22
- lsof -nPi:3306
- lsof /var/lib/mysql/mysql.sock

```
[wangqiang@dev242 ~]$ lsof -h
lsof 4.87
 latest revision: ftp://lsof.itap.purdue.edu/pub/tools/unix/lsof/
 latest FAQ: ftp://lsof.itap.purdue.edu/pub/tools/unix/lsof/FAQ
 latest man page: ftp://lsof.itap.purdue.edu/pub/tools/unix/lsof/lsof_man
 usage: [-?abhKlnNoOPRtUvVX] [+|-c c] [+|-d s] [+D D] [+|-f[gG]] [+|-e s]
  [-F [f]] [-g [s]] [-i [i]] [+|-L [l]] [+m [m]] [+|-M] [-o [o]] [-p s]
[+|-r [t]] [-s [p:s]] [-S [t]] [-T [t]] [-u s] [+|-w] [-x [fl]] [--] [names]
Defaults in parentheses; comma-separated set (s) items; dash-separated ranges.
  -?|-h list help          -a AND selections (OR)     -b avoid kernel blocks
  -c c  cmd c ^c /c/[bix]  +c w  COMMAND width (9)    +d s  dir s files
  -d s  select by FD set   +D D  dir D tree *SLOW?*   +|-e s  exempt s *RISKY*
  -i select IPv[46] files  -K list tasKs (threads)    -l list UID numbers
  -n no host names         -N select NFS files        -o list file offset
  -O no overhead *RISKY*   -P no port names           -R list paRent PID
  -s list file size        -t terse listing           -T disable TCP/TPI info
  -U select Unix socket    -v list version info       -V verbose search
  +|-w  warnings (+)       -X skip TCP&UDP* files     -Z Z  context [Z]
  -- end option scan
  +f|-f  +filesystem or -file names     +|-f[gG] flaGs
  -F [f] select fields; -F? for help
  +|-L [l] list (+) suppress (-) link counts < l (0 = all; default = 0)
                                        +m [m] use|create mount supplement
  +|-M   portMap registration (-)       -o o   o 0t offset digits (8)
  -p s   exclude(^)|select PIDs         -S [t] t second stat timeout (15)
  -T qs TCP/TPI Q,St (s) info
  -g [s] exclude(^)|select and print process group IDs
  -i i   select by IPv[46] address: [46][proto][@host|addr][:svc_list|port_list]
  +|-r [t[m<fmt>]] repeat every t seconds (15);  + until no files, - forever.
       An optional suffix to t is m<fmt>; m must separate t from <fmt> and
      <fmt> is an strftime(3) format for the marker line.
  -s p:s  exclude(^)|select protocol (p = TCP|UDP) states by name(s).
  -u s   exclude(^)|select login|UID set s
  -x [fl] cross over +d|+D File systems or symbolic Links
  names  select named files or files on named file systems
Anyone can list all files; /dev warnings disabled; kernel ID check disabled.
```

strace

- 跟踪某一进程执行的系统调用
- 性能分析，数据抓包，应用异常行为分析
- Attach到指定进程后，该进程性能会有约15%损失
- strace -f -s 1024 -ttT -p <pid> -o /tmp/***.log
- strace -f -c -p <pid>
- 关键字accept, connect, futex, read, write, send, recv…

```
[wangqiang@dev242 ~]$ strace -h
usage: strace [-CdffhiqrtttTvVxxy] [-I n] [-e expr]...
              [-a column] [-o file] [-s strsize] [-P path]...
              -p pid... / [-D] [-E var=val]... [-u username] PROG [ARGS]
   or: strace -c[df] [-I n] [-e expr]... [-O overhead] [-S sortby]
              -p pid... / [-D] [-E var=val]... [-u username] PROG [ARGS]
-c -- count time, calls, and errors for each syscall and report summary
-C -- like -c but also print regular output
-d -- enable debug output to stderr
-D -- run tracer process as a detached grandchild, not as parent
-f -- follow forks, -ff -- with output into separate files
-i -- print instruction pointer at time of syscall
-q -- suppress messages about attaching, detaching, etc.
-r -- print relative timestamp, -t -- absolute timestamp, -tt -- with usecs
-T -- print time spent in each syscall
-v -- verbose mode: print unabbreviated argv, stat, termios, etc. args
-x -- print non-ascii strings in hex, -xx -- print all strings in hex
-y -- print paths associated with file descriptor arguments
-h -- print help message, -V -- print version
-a column -- alignment COLUMN for printing syscall results (default 40)
-b execve -- detach on this syscall
-e expr -- a qualifying expression: option=[!]all or option=[!]val1[,val2]...
   options: trace, abbrev, verbose, raw, signal, read, write
-I interruptible --
   1: no signals are blocked
   2: fatal signals are blocked while decoding syscall (default)
   3: fatal signals are always blocked (default if '-o FILE PROG')
   4: fatal signals and SIGTSTP (^Z) are always blocked
      (useful to make 'strace -o FILE PROG' not stop on ^Z)
-o file -- send trace output to FILE instead of stderr
-O overhead -- set overhead for tracing syscalls to OVERHEAD usecs
-p pid -- trace process with process id PID, may be repeated
-s strsize -- limit length of print strings to STRSIZE chars (default 32)
-S sortby -- sort syscall counts by: time, calls, name, nothing (default time)
-u username -- run command as username handling setuid and/or setgid
-E var=val -- put var=val in the environment for command
-E var -- remove var from the environment for command
-P path -- trace accesses to path
```

ss

- 查看TCP/UDP sockets状态
- ss -an | grep LISTEN
- ss -s
- netstat -s
- EST, CLOSE-WAIT, TIME-WAIT, SYN_SENT
- EPIPE，Broken pipe
- 发送TCP RST可以避免进入TIME-WAIT状态

```
[wangqiang@dev242 ~]$ ss -h
Usage: ss [ OPTIONS ]
       ss [ OPTIONS ] [ FILTER ]
   -h, --help          this message
   -V, --version       output version information
   -n, --numeric       don't resolve service names
   -r, --resolve       resolve host names
   -a, --all           display all sockets
   -l, --listening     display listening sockets
   -o, --options       show timer information
   -e, --extended      show detailed socket information
   -m, --memory        show socket memory usage
   -p, --processes     show process using socket
   -i, --info          show internal TCP information
   -s, --summary       show socket usage summary
   -b, --bpf           show bpf filter socket information
   -Z, --context       display process SELinux security contexts
   -z, --contexts      display process and socket SELinux security contexts
   -N, --net           switch to the specified network namespace name

   -4, --ipv4          display only IP version 4 sockets
   -6, --ipv6          display only IP version 6 sockets
   -0, --packet        display PACKET sockets
   -t, --tcp           display only TCP sockets
   -u, --udp           display only UDP sockets
   -d, --dccp          display only DCCP sockets
   -w, --raw           display only RAW sockets
   -x, --unix          display only Unix domain sockets
   -f, --family=FAMILY display sockets of type FAMILY

   -A, --query=QUERY, --socket=QUERY
       QUERY := {all|inet|tcp|udp|raw|unix|unix_dgram|unix_stream|unix_seqpacket|packet|
netlink}[,QUERY]

   -D, --diag=FILE     Dump raw information about TCP sockets to FILE
   -F, --filter=FILE   read filter information from FILE
       FILTER := [ state STATE-FILTER ] [ EXPRESSION ]
       STATE-FILTER := {all|connected|synchronized|bucket|big|TCP-STATES}
         TCP-STATES := {established|syn-sent|syn-recv|fin-wait-{1,2}|time-wait|closed|cl
ose-wait|last-ack|listen|closing}
          connected := {established|syn-sent|syn-recv|fin-wait-{1,2}|time-wait|close-wai
t|last-ack|closing}
       synchronized := {established|syn-recv|fin-wait-{1,2}|time-wait|close-wait|last-ac
k|closing}
             bucket := {syn-recv|time-wait}
                big := {established|syn-sent|fin-wait-{1,2}|closed|close-wait|last-ack|l
isten|closing}
```

tcpdump

- 网络抓包，网络故障分析
- tcpdump -vv -i eth0 port 3306 -X -s 0
- tcpdump -vv -s 0 -i eth0 port 80 -w /tmp/tcp80.pcap
- tcpdump -vv -s 0 -i eth0 port 9527 and host 54.222.206.6 -w /tmp/54_9527.pcap
- 快速查看包ASCII内容：strings ***.pcap | less

```
[wangqiang@dev242 ~]$ tcpdump -h
tcpdump version 4.5.1
libpcap version 1.5.3
Usage: tcpdump [-aAbdDefhHIJKlLnNOpqRStuUvxX] [ -B size ] [ -c count ]
                [ -C file_size ] [ -E algo:secret ] [ -F file ] [ -G seconds ]
                [ -i interface ] [ -j tstamptype ] [ -M secret ]
                [ -P in|out|inout ]
                [ -r file ] [ -s snaplen ] [ -T type ] [ -V file ] [ -w file ]
                [ -W filecount ] [ -y datalinktype ] [ -z command ]
                [ -Z user ] [ expression ]
```

objdump

- 查看二进制文件对应的反汇编代码
- objdump -D /usr/local/webserver/erlang/lib/erlang/erts-8.2/bin/beam.smp > /tmp/beam.smp.objdump
- 查看崩溃指令地址（IP）对应的函数

```
[wangqiang@dev242 ~]$ objdump --help
Usage: objdump <option(s)> <file(s)>
 Display information from object <file(s)>.
 At least one of the following switches must be given:
  -a, --archive-headers    Display archive header information
  -f, --file-headers       Display the contents of the overall file header
  -p, --private-headers    Display object format specific file header contents
  -P, --private=OPT,OPT... Display object format specific contents
  -h, --[section-]headers  Display the contents of the section headers
  -x, --all-headers        Display the contents of all headers
  -d, --disassemble        Display assembler contents of executable sections
  -D, --disassemble-all    Display assembler contents of all sections
  -S, --source             Intermix source code with disassembly
  -s, --full-contents      Display the full contents of all sections requested
  -g, --debugging          Display debug information in object file
  -e, --debugging-tags     Display debug information using ctags style
  -G, --stabs              Display (in raw form) any STABS info in the file
  -W[lLiaprmfFsoRt] or
  --dwarf[=rawline,=decodedline,=info,=abbrev,=pubnames,=aranges,=macro,=frames,
          =frames-interp,=str,=loc,=Ranges,=pubtypes,
          =gdb_index,=trace_info,=trace_abbrev,=trace_aranges,
          =addr,=cu_index]
                           Display DWARF info in the file
  -t, --syms               Display the contents of the symbol table(s)
  -T, --dynamic-syms       Display the contents of the dynamic symbol table
  -r, --reloc              Display the relocation entries in the file
  -R, --dynamic-reloc      Display the dynamic relocation entries in the file
  @<file>                  Read options from <file>
  -v, --version            Display this program's version number
  -i, --info               List object formats and architectures supported
  -H, --help               Display this information

 The following switches are optional:
  -b, --target=BFDNAME           Specify the target object format as BFDNAME
  -m, --architecture=MACHINE     Specify the target architecture as MACHINE
  -j, --section=NAME             Only display information for section NAME
  -M, --disassembler-options=OPT Pass text OPT on to the disassembler
  -EB --endian=big               Assume big endian format when disassembling
  -EL --endian=little            Assume little endian format when disassembling
```

gcore/gdb

- 调试进程&core文件，模拟慢响应
- 模拟进程崩溃：kill -<SIGNAL> <pid>
- 生成指定进程的core文件：gcore <pid>
- 在线调试进程：gdb -p <pid>
- 调试子进程：(gdb) set follow-fork-mode child
- 调试多线程：(gdb) info thr , thr <thr_id>
- 断点技巧：(gdb) b read, write, accept, 业务函数
- gcore对占用过高虚拟内存（32G+）的进程，应慎重

```
[wangqiang@dev242 ~]$ gdb -h
This is the GNU debugger.  Usage:

    gdb [options] [executable-file [core-file or process-id]]
    gdb [options] --args executable-file [inferior-arguments ...]
    gdb [options] [--python|-P] script-file [script-arguments ...]

Options:

  --args             Arguments after executable-file are passed to inferior
  -b BAUDRATE        Set serial port baud rate used for remote debugging.
  --batch            Exit after processing options.
  --batch-silent     As for --batch, but suppress all gdb stdout output.
  --return-child-result
                     GDB exit code will be the child's exit code.
  --cd=DIR           Change current directory to DIR.
  --command=FILE, -x Execute GDB commands from FILE.
  --eval-command=COMMAND, -ex
                     Execute a single GDB command.
                     May be used multiple times and in conjunction
                     with --command.
  --init-command=FILE, -ix Like -x but execute it before loading inferior.
  --init-eval-command=COMMAND, -iex Like -ex but before loading inferior.
  --core=COREFILE    Analyze the core dump COREFILE.
  --pid=PID          Attach to running process PID.
  --dbx              DBX compatibility mode.
  --directory=DIR    Search for source files in DIR.
  --exec=EXECFILE    Use EXECFILE as the executable.
  --fullname         Output information used by emacs-GDB interface.
  --help             Print this message.
  --interpreter=INTERP
                     Select a specific interpreter / user interface
  -l TIMEOUT         Set timeout in seconds for remote debugging.
  --nw               Do not use a window interface.
  --nx               Do not read any .gdbinit files.
  --nh               Do not read .gdbinit file from home directory.
  --python, -P       Following argument is Python script file; remaining
```

线上问题分析举例

- **K8s http svc不稳定**：tcpdump抓包发现多个7层负载端IP在容器宿主机节点被SNAT映射为同一源IP
- **Go容器优化**：容器内设置GOMAXPROCS，防止过多的工作线程争用有限的容器资源CPU，如：高iowait
- **Node慢响应**：strace抓取相关node进程，向mysql发送请求后未调用接收应答--node连接池使用BUG
- **Rabbitmq崩溃**：ERL-486，根据系统messages的指令地址结合objdump查找对应的反汇编函数
- **Rabbitmq满载**：iftop查看MQ最大流量IP，ss查看该客户端IP端口，lsof根据端口号定位至进程，strace抓取该进程运行细节—PHP进程向MQ pub大数据分配内存失败崩溃后不断重启重试。
- **DNS解析开销**：A/PTR，使用strace分析客户端或服务端进程、或使用tcpdump分析
- **应用调用链分析**：使用lsof及strace命令分析应用进程与其他组件调用关系，数据请求及应答数据

Thanks !