The Performance Debugging Toolkit

Under the hood of Linux systems

So you have your project

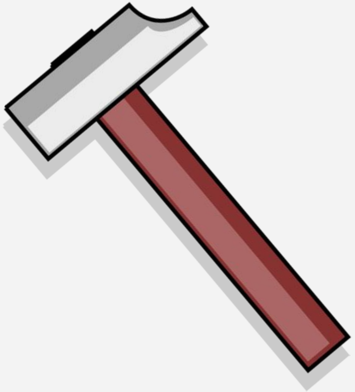
It should work

But…

It should work, but...

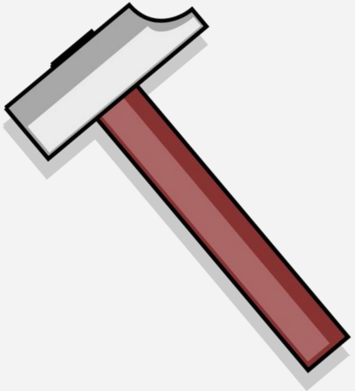
- Performance is not as expected
- Unpredictable results
- Unexpected resource utilization
- Weird system behavior

What to do???



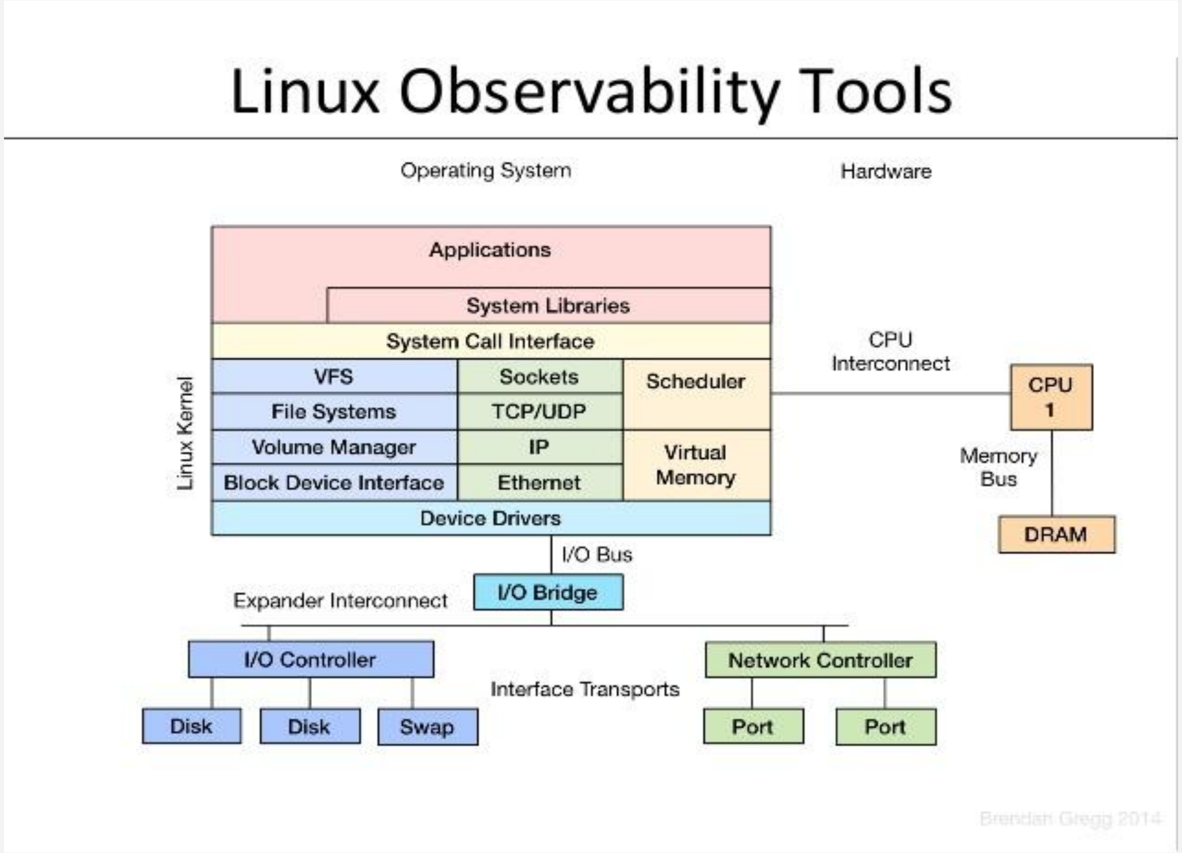
- Add prints
- Time execution
- Random parameter tuning
- Improvised optimizations
- Execute until a good measurement happens?

What to do???



- ~~Add prints~~
- ~~Time execution~~
- ~~Random parameter tuning~~
- ~~Improvised optimizations~~
- ~~Execute until a good measurement happens?~~

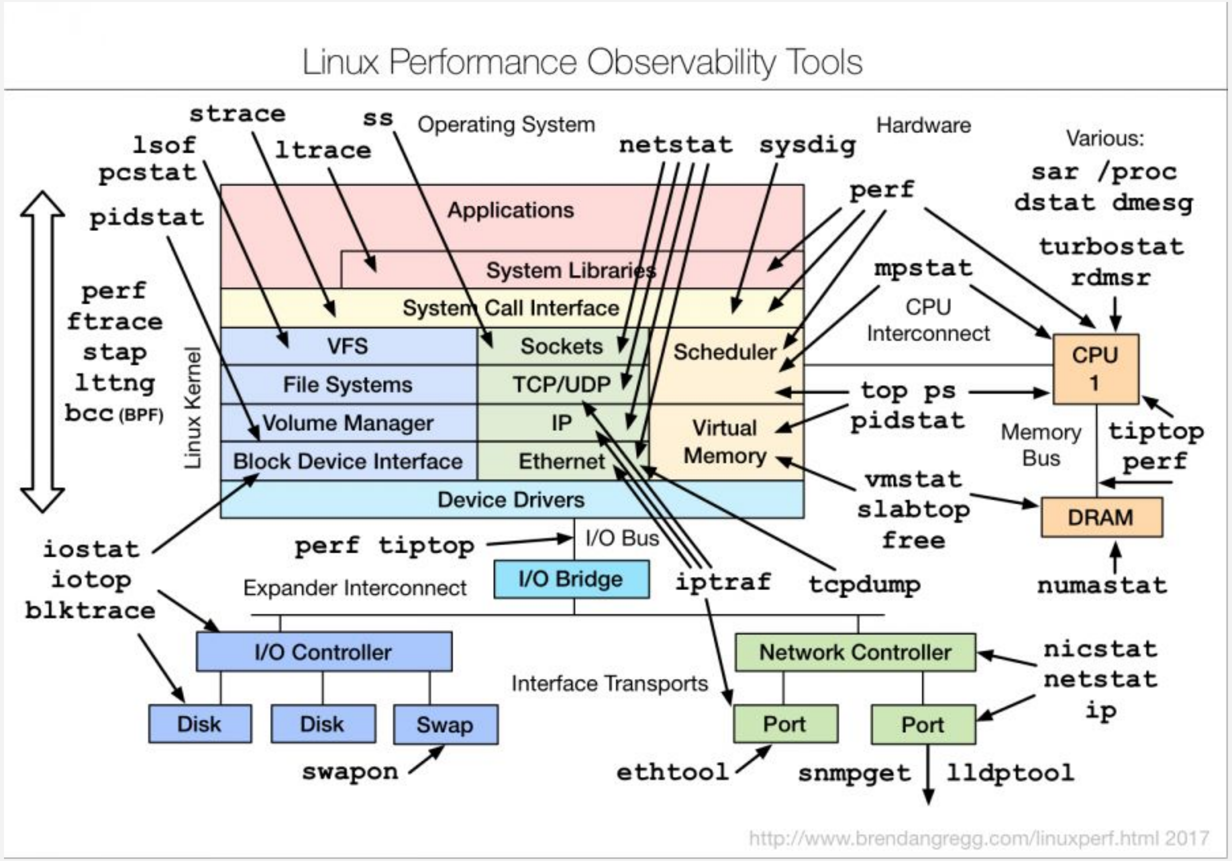
Where to start?



Linux Observability Tools

Brendan Gregg 2014

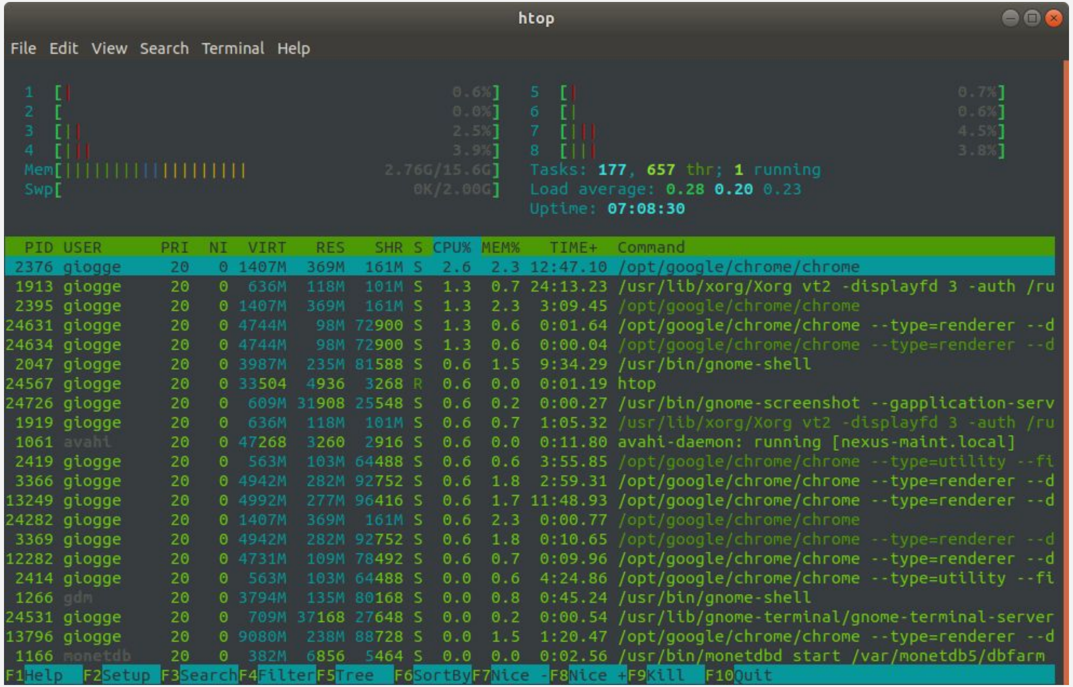
Well... get the toolbox!



Linux Performance Observability Tools

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

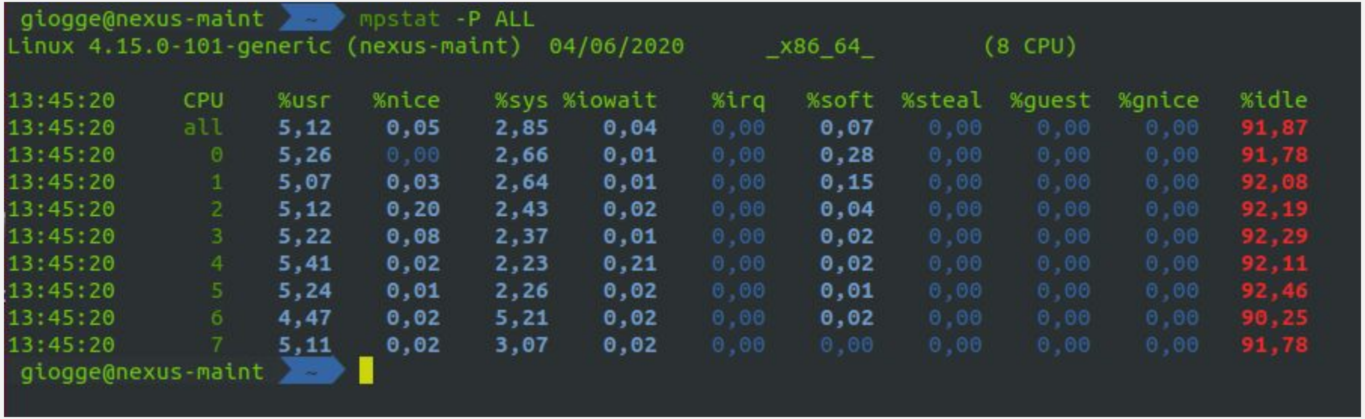
CPU - top, htop, ps



```
  1  [|                                  0.6%]   5  [|                                   0.7%]
  2  [                                   0.0%]   6  [|                                   0.6%]
  3  [||                                 2.5%]   7  [|||                                 4.5%]
  4  [|||                                3.9%]   8  [|||                                 3.8%]
  Mem[|||||||||||||||||||||      2.76G/15.6G]   Tasks: 177, 657 thr; 1 running
  Swp[                               0K/2.00G]   Load average: 0.28 0.20 0.23
                                                 Uptime: 07:08:30

  PID USER      PRI  NI  VIRT   RES   SHR S CPU% MEM%   TIME+  Command
 2376 giogge     20   0 1407M  369M  161M S  2.6  2.3 12:47.10 /opt/google/chrome/chrome
 1913 giogge     20   0  636M  118M  101M S  1.3  0.7 24:13.23 /usr/lib/xorg/Xorg vt2 -displayfd 3 -auth /ru
 2395 giogge     20   0 1407M  369M  161M S  1.3  2.3  3:09.45 /opt/google/chrome/chrome
24631 giogge     20   0 4744M   98M 72900 S  1.3  0.6  0:01.64 /opt/google/chrome/chrome --type=renderer --d
24634 giogge     20   0 4744M   98M 72900 S  1.3  0.6  0:00.04 /opt/google/chrome/chrome --type=renderer --d
 2047 giogge     20   0 3987M  235M 81588 S  0.6  1.5  9:34.29 /usr/bin/gnome-shell
24567 giogge     20   0 33504  4936  3268 R  0.6  0.0  0:01.19 htop
24726 giogge     20   0  609M 31908 25548 S  0.6  0.2  0:00.27 /usr/bin/gnome-screenshot --gapplication-serv
 1919 giogge     20   0  636M  118M  101M S  0.6  0.7  1:05.32 /usr/lib/xorg/Xorg vt2 -displayfd 3 -auth /ru
 1061 avahi      20   0 47268  3260  2916 S  0.6  0.0  0:11.80 avahi-daemon: running [nexus-maint.local]
 2419 giogge     20   0  563M  103M 64488 S  0.6  0.6  3:55.85 /opt/google/chrome/chrome --type=utility --fi
 3366 giogge     20   0 4942M  282M 92752 S  0.6  1.8  2:59.31 /opt/google/chrome/chrome --type=renderer --d
13249 giogge     20   0 4992M  277M 96416 S  0.6  1.7 11:48.93 /opt/google/chrome/chrome --type=renderer --d
24282 giogge     20   0 1407M  369M  161M S  0.6  2.3  0:00.77 /opt/google/chrome/chrome
 3369 giogge     20   0 4942M  282M 92752 S  0.6  1.8  0:10.65 /opt/google/chrome/chrome --type=renderer --d
12282 giogge     20   0 4731M  109M 78492 S  0.6  0.7  0:09.96 /opt/google/chrome/chrome --type=renderer --d
 2414 giogge     20   0  563M  103M 64488 S  0.0  0.6  4:24.86 /opt/google/chrome/chrome --type=utility --fi
 1266 gdm        20   0 3794M  135M 80168 S  0.0  0.8  0:45.24 /usr/bin/gnome-shell
24531 giogge     20   0  709M 37168 27648 S  0.0  0.2  0:00.54 /usr/lib/gnome-terminal/gnome-terminal-server
13796 giogge     20   0 9080M  238M 88728 S  0.0  1.5  1:20.47 /opt/google/chrome/chrome --type=renderer --d
 1166 monetdb    20   0  382M  6856  5464 S  0.0  0.0  0:02.56 /usr/bin/monetdbd start /var/monetdb5/dbfarm
F1Help  F2Setup F3SearchF4FilterF5Tree  F6SortByF7Nice -F8Nice +F9Kill  F10Quit
```

https://linux.die.net/man/1/top https://linux.die.net/man/1/htop
https://www.man7.org/linux/man-pages/man1/ps.1.htm

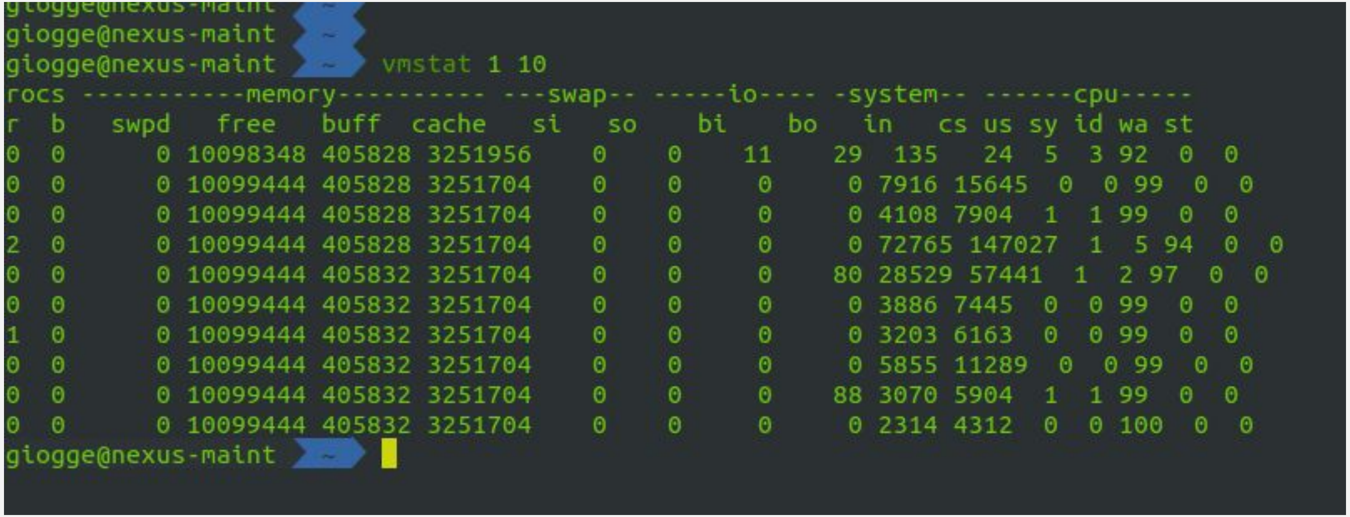
Multicore - mpstat



```
 giogge@nexus-maint  ~  mpstat -P ALL
Linux 4.15.0-101-generic (nexus-maint)  04/06/2020      _x86_64_        (8 CPU)

13:45:20     CPU    %usr   %nice    %sys %iowait    %irq   %soft  %steal  %guest  %gnice   %idle
13:45:20     all    5,12    0,05    2,85    0,04    0,00    0,07    0,00    0,00    0,00   91,87
13:45:20       0    5,26    0,00    2,66    0,01    0,00    0,28    0,00    0,00    0,00   91,78
13:45:20       1    5,07    0,03    2,64    0,01    0,00    0,15    0,00    0,00    0,00   92,08
13:45:20       2    5,12    0,20    2,43    0,02    0,00    0,04    0,00    0,00    0,00   92,19
13:45:20       3    5,22    0,08    2,37    0,01    0,00    0,02    0,00    0,00    0,00   92,29
13:45:20       4    5,41    0,02    2,23    0,21    0,00    0,02    0,00    0,00    0,00   92,11
13:45:20       5    5,24    0,01    2,26    0,02    0,00    0,01    0,00    0,00    0,00   92,46
13:45:20       6    4,47    0,02    5,21    0,02    0,00    0,02    0,00    0,00    0,00   90,25
13:45:20       7    5,11    0,02    3,07    0,02    0,00    0,00    0,00    0,00    0,00   91,78
 giogge@nexus-maint  ~
```

https://linux.die.net/man/1/mpstat

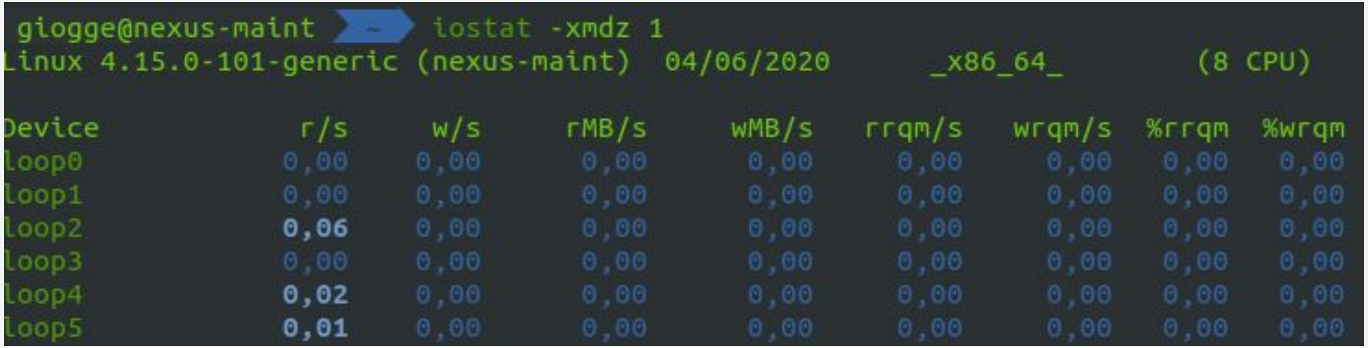
Memory - vmstat



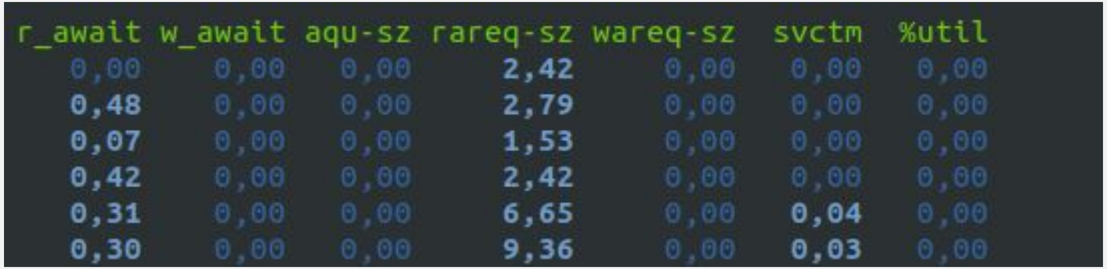
```
giogge@nexus-maint  ~
giogge@nexus-maint  ~
giogge@nexus-maint  ~  vmstat 1 10
rocs -----------memory---------- ---swap-- -----io---- -system-- ------cpu-----
r  b   swpd   free   buff  cache   si   so    bi    bo   in   cs us sy id wa st
0  0      0 10098348 405828 3251956    0    0    11    29  135   24  5  3 92  0  0
0  0      0 10099444 405828 3251704    0    0     0     0 7916 15645  0  0 99  0  0
0  0      0 10099444 405828 3251704    0    0     0     0 4108 7904  1  1 99  0  0
2  0      0 10099444 405828 3251704    0    0     0     0 72765 147027  1  5 94  0  0
0  0      0 10099444 405832 3251704    0    0     0    80 28529 57441  1  2 97  0  0
0  0      0 10099444 405832 3251704    0    0     0     0 3886 7445  0  0 99  0  0
1  0      0 10099444 405832 3251704    0    0     0     0 3203 6163  0  0 99  0  0
0  0      0 10099444 405832 3251704    0    0     0     0 5855 11289  0  0 99  0  0
0  0      0 10099444 405832 3251704    0    0     0    88 3070 5904  1  1 99  0  0
0  0      0 10099444 405832 3251704    0    0     0     0 2314 4312  0  0 100  0  0
giogge@nexus-maint  ~
```

https://linux.die.net/man/8/vmstat

I/O - iostat



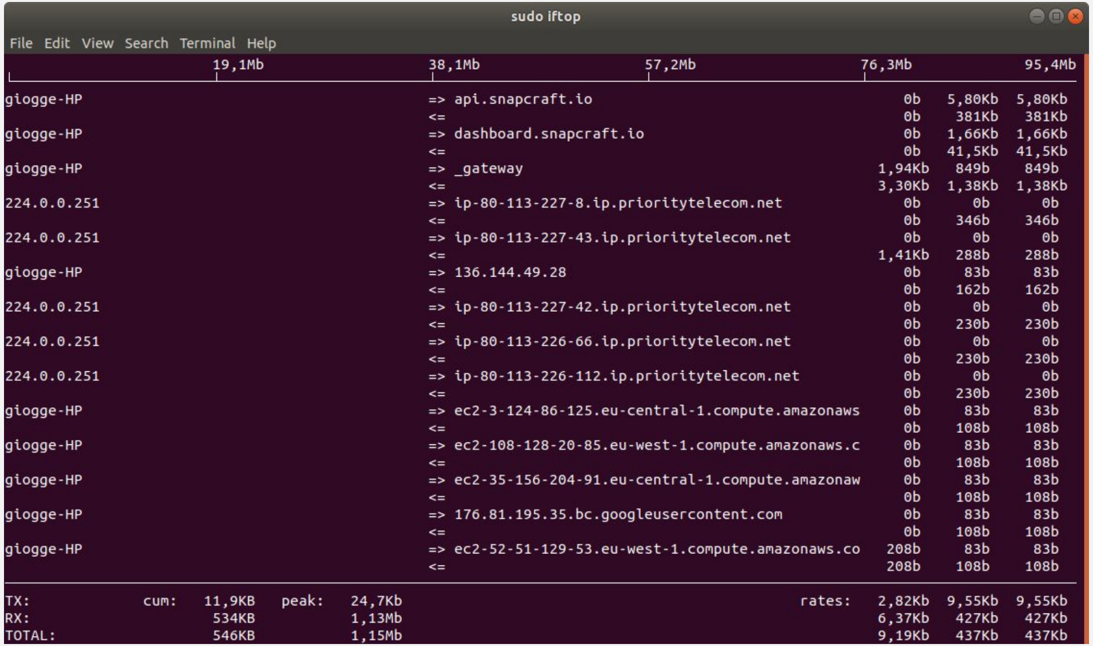
```
giogge@nexus-maint  ~  iostat -xmdz 1
Linux 4.15.0-101-generic (nexus-maint)  04/06/2020      _x86_64_       (8 CPU)

Device            r/s     w/s     rMB/s     wMB/s    rrqm/s    wrqm/s  %rrqm  %wrqm
loop0            0,00    0,00      0,00      0,00      0,00      0,00   0,00   0,00
loop1            0,00    0,00      0,00      0,00      0,00      0,00   0,00   0,00
loop2            0,06    0,00      0,00      0,00      0,00      0,00   0,00   0,00
loop3            0,00    0,00      0,00      0,00      0,00      0,00   0,00   0,00
loop4            0,02    0,00      0,00      0,00      0,00      0,00   0,00   0,00
loop5            0,01    0,00      0,00      0,00      0,00      0,00   0,00   0,00
```



```
r_await w_await aqu-sz rareq-sz wareq-sz  svctm  %util
   0,00    0,00   0,00     2,42     0,00   0,00   0,00
   0,48    0,00   0,00     2,79     0,00   0,00   0,00
   0,07    0,00   0,00     1,53     0,00   0,00   0,00
   0,42    0,00   0,00     2,42     0,00   0,00   0,00
   0,31    0,00   0,00     6,65     0,00   0,04   0,00
   0,30    0,00   0,00     9,36     0,00   0,03   0,00
```

https://linux.die.net/man/1/iostat

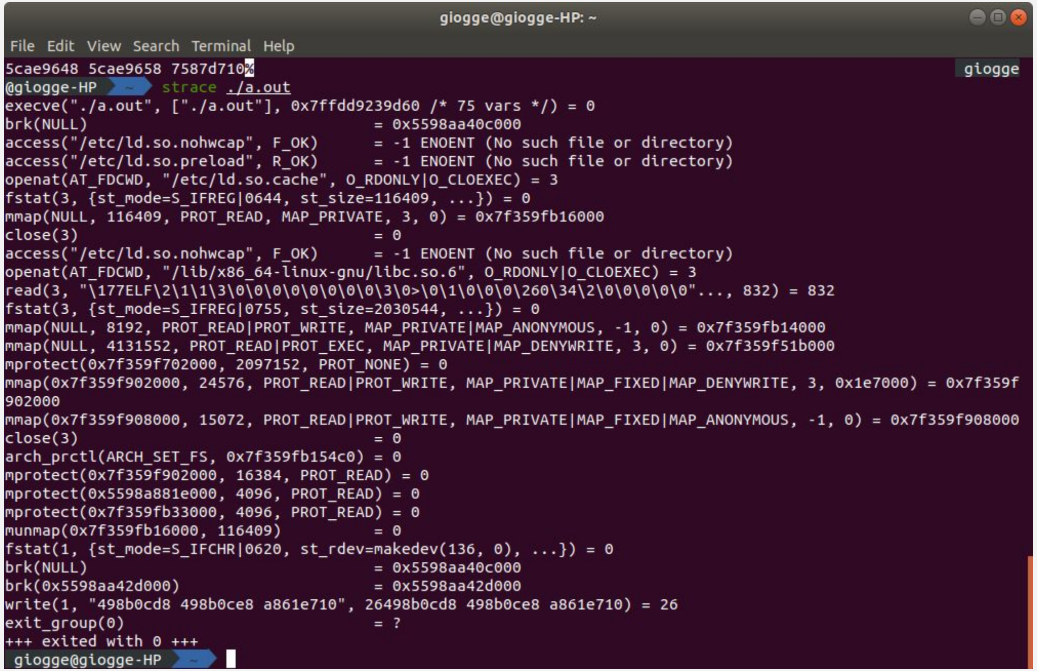
Networking - iftop, netstat



```
                       19,1Mb                  38,1Mb                  57,2Mb                  76,3Mb           95,4Mb
└───────────────────────┴───────────────────────┴───────────────────────┴───────────────────────┴──────────────────
giogge-HP                                       => api.snapcraft.io                                 0b  5,80Kb  5,80Kb
                                                <=                                                  0b   381Kb   381Kb
giogge-HP                                       => dashboard.snapcraft.io                           0b  1,66Kb  1,66Kb
                                                <=                                                  0b  41,5Kb  41,5Kb
giogge-HP                                       => _gateway                                     1,94Kb    849b    849b
                                                <=                                              3,30Kb  1,38Kb  1,38Kb
224.0.0.251                                     => ip-80-113-227-8.ip.prioritytelecom.net           0b      0b      0b
                                                <=                                                  0b    346b    346b
224.0.0.251                                     => ip-80-113-227-43.ip.prioritytelecom.net          0b      0b      0b
                                                <=                                              1,41Kb    288b    288b
giogge-HP                                       => 136.144.49.28                                    0b     83b     83b
                                                <=                                                  0b    162b    162b
224.0.0.251                                     => ip-80-113-227-42.ip.prioritytelecom.net          0b      0b      0b
                                                <=                                                  0b    230b    230b
224.0.0.251                                     => ip-80-113-226-66.ip.prioritytelecom.net          0b      0b      0b
                                                <=                                                  0b    230b    230b
224.0.0.251                                     => ip-80-113-226-112.ip.prioritytelecom.net         0b      0b      0b
                                                <=                                                  0b    230b    230b
giogge-HP                                       => ec2-3-124-86-125.eu-central-1.compute.amazonaws  0b     83b     83b
                                                <=                                                  0b    108b    108b
giogge-HP                                       => ec2-108-128-20-85.eu-west-1.compute.amazonaws.c  0b     83b     83b
                                                <=                                                  0b    108b    108b
giogge-HP                                       => ec2-35-156-204-91.eu-central-1.compute.amazonaw  0b     83b     83b
                                                <=                                                  0b    108b    108b
giogge-HP                                       => 176.81.195.35.bc.googleusercontent.com           0b     83b     83b
                                                <=                                                  0b    108b    108b
giogge-HP                                       => ec2-52-51-129-53.eu-west-1.compute.amazonaws.co 208b    83b     83b
                                                <=                                                 208b   108b    108b
──────────────────────────────────────────────────────────────────────────────────────────────────────────────────
TX:             cum:   11,9KB   peak:   24,7Kb                                         rates:   2,82Kb  9,55Kb  9,55Kb
RX:                     534KB           1,13Mb                                                  6,37Kb   427Kb   427Kb
TOTAL:                  546KB           1,15Mb                                                  9,19Kb   437Kb   437Kb
```

https://linux.die.net/man/8/iftop https://linux.die.net/man/8/netstat

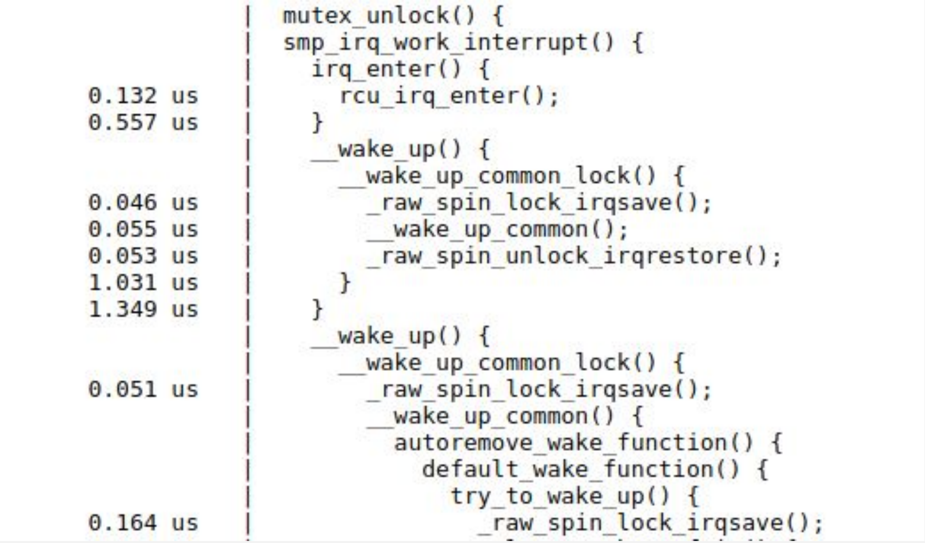
Execution - ltrace, strace



```
5cae9648 5cae9658 7587d710%                                                                                  giogge
@giogge-HP  ~  strace ./a.out
execve("./a.out", ["./a.out"], 0x7ffdd9239d60 /* 75 vars */) = 0
brk(NULL)                               = 0x5598aa40c000
access("/etc/ld.so.nohwcap", F_OK)      = -1 ENOENT (No such file or directory)
access("/etc/ld.so.preload", R_OK)      = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/etc/ld.so.cache", O_RDONLY|O_CLOEXEC) = 3
fstat(3, {st_mode=S_IFREG|0644, st_size=116409, ...}) = 0
mmap(NULL, 116409, PROT_READ, MAP_PRIVATE, 3, 0) = 0x7f359fb16000
close(3)                                = 0
access("/etc/ld.so.nohwcap", F_OK)      = -1 ENOENT (No such file or directory)
openat(AT_FDCWD, "/lib/x86_64-linux-gnu/libc.so.6", O_RDONLY|O_CLOEXEC) = 3
read(3, "\177ELF\2\1\1\3\0\0\0\0\0\0\0\0\3\0>\0\1\0\0\0\260\34\2\0\0\0\0\0"..., 832) = 832
fstat(3, {st_mode=S_IFREG|0755, st_size=2030544, ...}) = 0
mmap(NULL, 8192, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) = 0x7f359fb14000
mmap(NULL, 4131552, PROT_READ|PROT_EXEC, MAP_PRIVATE|MAP_DENYWRITE, 3, 0) = 0x7f359f51b000
mprotect(0x7f359f702000, 2097152, PROT_NONE) = 0
mmap(0x7f359f902000, 24576, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_FIXED|MAP_DENYWRITE, 3, 0x1e7000) = 0x7f359f
902000
mmap(0x7f359f908000, 15072, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_FIXED|MAP_ANONYMOUS, -1, 0) = 0x7f359f908000
close(3)                                = 0
arch_prctl(ARCH_SET_FS, 0x7f359fb154c0) = 0
mprotect(0x7f359f902000, 16384, PROT_READ) = 0
mprotect(0x5598a881e000, 4096, PROT_READ) = 0
mprotect(0x7f359fb33000, 4096, PROT_READ) = 0
munmap(0x7f359fb16000, 116409)          = 0
fstat(1, {st_mode=S_IFCHR|0620, st_rdev=makedev(136, 0), ...}) = 0
brk(NULL)                               = 0x5598aa40c000
brk(0x5598aa42d000)                     = 0x5598aa42d000
write(1, "498b0cd8 498b0ce8 a861e710", 26498b0cd8 498b0ce8 a861e710) = 26
exit_group(0)                           = ?
+++ exited with 0 +++
 giogge@giogge-HP  ~ 
```

https://www.man7.org/linux/man-pages/man1/strace.1.html https://www.man7.org/linux/man-pages/man1/ltrace.1.html

Kernel calls - ftrace



```
                  |  mutex_unlock() {
                  |  smp_irq_work_interrupt() {
                  |    irq_enter() {
      0.132 us    |      rcu_irq_enter();
      0.557 us    |    }
                  |    __wake_up() {
                  |      __wake_up_common_lock() {
      0.046 us    |        _raw_spin_lock_irqsave();
      0.055 us    |        __wake_up_common();
      0.053 us    |        _raw_spin_unlock_irqrestore();
      1.031 us    |      }
      1.349 us    |    }
                  |    __wake_up() {
                  |      __wake_up_common_lock() {
      0.051 us    |        _raw_spin_lock_irqsave();
                  |        __wake_up_common() {
                  |          autoremove_wake_function() {
                  |            default_wake_function() {
                  |              try_to_wake_up() {
      0.164 us    |                _raw_spin_lock_irqsave();
```

https://jvns.ca/blog/2017/03/19/getting-started-with-ftrace/

/proc - all things counters

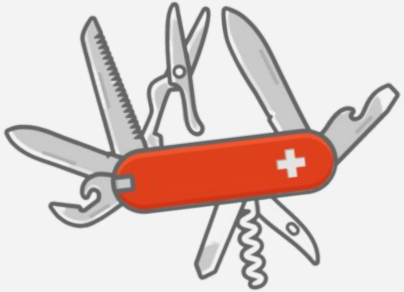
- A filesystem for system information
- The tools probably read this
- You could write your own!

- /proc/cpuinfo
- /proc/devices
- /proc/ksyms

https://www.tldp.org/LDP/sag/html/proc-fs.html

The (un)biased favourite

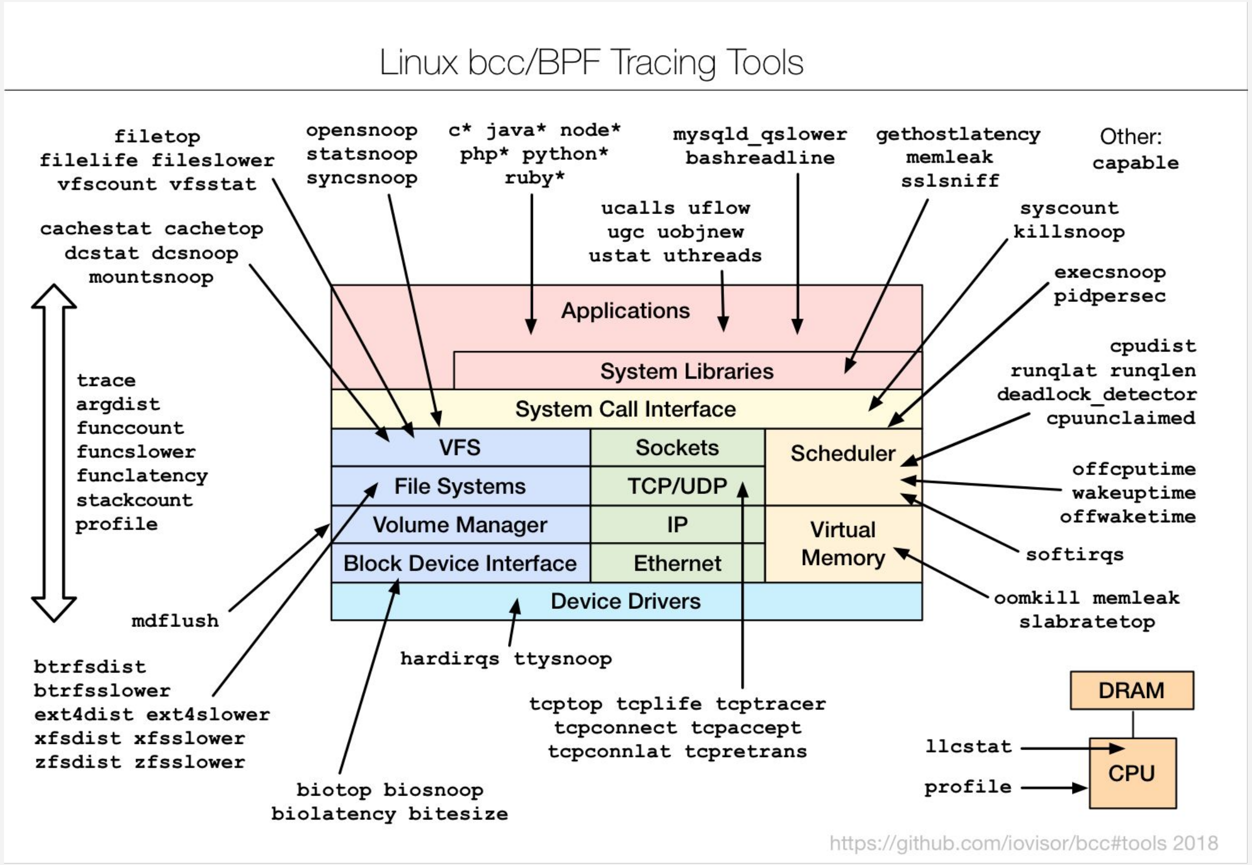
eBPF - the swiss army knife



- Basically all we just said
- … on steroids!
- Easy to use: BCC
- Not for the faint of heart: native

https://github.com/iovisor/bcc

One tool to rule them all



Linux bcc/BPF Tracing Tools

https://github.com/iovisor/bcc#tools 2018

Distributed systems

- Look into debug tools in frameworks
- Look at your systems capability (e.g. hadoop)
- Distributed monitoring (e.g. ganglia)

http://ganglia.info/

Now we have the tools, but we still need the manual

Remember!

- Run long enough
- Measure one level deeper
- Crosscheck with more subsystems
- Don’t trust your numbers
- Back-of-the-envelope estimates

Minimize your entropy

- Execute in controlled environment
- Reduce randomness
- Cgroups
- Cpulimit
- Numactl

https://www.man7.org/linux/man-pages/man7/cgroups.7.html https://linux.die.net/man/8/numactl
http://manpages.ubuntu.com/manpages/xenial/man1/cpulimit.1.html

Minimize your entropy

- Execute in controlled environment
- Reduce randomness
- Cgroups
- Cpulimit
- Numactl

DIY CONTAINERS!!!

https://www.man7.org/linux/man-pages/man7/cgroups.7.html https://linux.die.net/man/8/numactl
http://manpages.ubuntu.com/manpages/xenial/man1/cpulimit.1.html

You don’t have a hammer, don’t treat everything as a nail!

References

- http://www.brendangregg.com/Slides/Velocity2015_LinuxPerfTools.pdf
- http://www.brendangregg.com/blog/2019-01-01/learn-ebpf-tracing.html
- https://cacm.acm.org/magazines/2018/7/229031-always-measure-one-level-deeper/fulltext
- https://github.com/iovisor/bcc
- https://lwn.net/Articles/365835/