Introduction to Linux System Performance Analysis

Suresh Kumar Ponnusamy

September 27, 2017

Outline

1. System
2. Application
3. Profilers
4. Debuggers
5. Books

Topic

1. System
2. Application
3. Profilers
4. Debuggers
5. Books

System performance analysis I

- USE (Utilization Saturation and Errors) Method [1]
- Identify and list the important resources (either physical or software)
  - Example - Hardware: CPU, Memory, Disk, Network etc
  - Example - Software: Locks, processes/threads capacity, file descriptor limit etc
- For each resource
  - Check utilization: How much is it being utilized? Is current utilization "safe"? or Could it lead to problems? It depends on the "resource"
  - Check saturation: Is the resource completely utilized and has extra work queued up? How much is the wait time? Is it within the acceptable range or could it lead to problems?
  - Check errors: Is it malfunctioning? generating some errors?
- A collection of this list + utilities is listed here [2] , [3], including a picture from there:

System performance analysis II

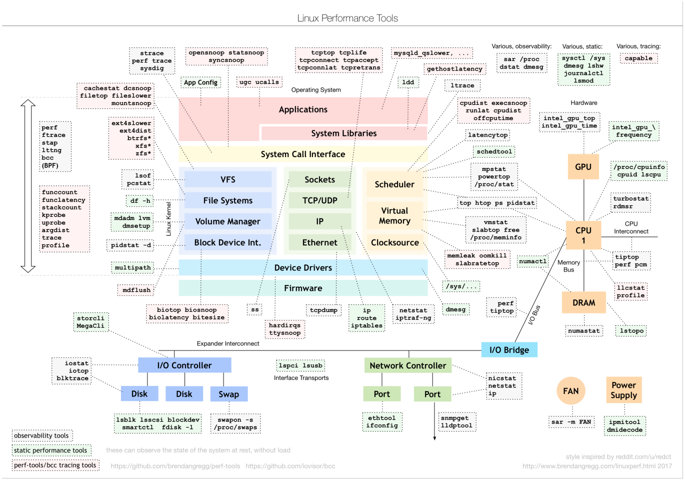


Figure: Linux Performance Tools

[1] http://www.brendangregg.com/usemethod.html

[2] http://www.brendangregg.com/USEmethod/use-linux.html

Resource: CPU I

- uptime/loadavg, top/htop/atop, vmstat, dmstat, "perf sched latency" etc
- Notes
  - loadavg not only shows CPU utilization but also includes other "resources" (i.e., any process that is in UNINTERRUPTIBLE state is also included: primary example is waiting for Disk IO but also includes others like certain locks). For more detailed analysis, you can see here [4]
  - Remember that CPU usage also includes waiting for memory access (Example: trying to access a memory location that is not in cache, it is being brought from RAM, the whole time, CPU will be spinning, waiting for the data)
  - Impact of virtualization: Keep an eye on "steal%"
    - Underlying physical hardware is shared with other workloads from "unknown" individuals/companies
    - Depending on what kind of work it is being done in other instance(s), our performance may be impacted (for example: other instance(s) may be saturating memory bandwidth etc)

Resource: CPU II

```
# uptime
uptime
 11:42:54 up 8 days,  1:09,  1 user,  load average: 0.24, 0.43, 0.48

############################

# Processes that are in running + uninterruptible state
# ps axl | awk '$10 ~ /[DR]/'
1     0    79     2  20   0      0     0 -      R    ?          0:24 [kswapd0]
4     0 24960 24959  20   0  14384  9824 -      D+   pts/22     0:07 dd if=/dev/sda of=/dev/null bs=8M count=10
0  1000 25359  5385  20   0  27636  2800 -      R+   pts/17     0:00 ps axl


############################

# vmstat 1
procs -----------memory---------- ---swap-- -----io---- -system-- ------cpu-----
 r  b   swpd   free   buff  cache   si   so    bi    bo   in   cs us sy id wa st
 1  0      0 2002984 748396 8277112    0    0     8    59   33   43  9  4 87  0  0
 0  0      0 1992204 748396 8287912    0    0     0   148 1096 3697  5  1 94  0  0

############################

# dstat -c 1
--total-cpu-usage--
usr sys idl wai stl
  9   4  87   0   0
  1   1  98   0   0
  3   1  96   0   0
  2   1  98   0   0
```

Resource: CPU III

```
  1   1  98   0   0

############################

# top
top - 11:53:53 up 8 days,  1:20,  1 user,  load average: 1.81, 0.96, 0.65
Tasks: 296 total,   1 running, 295 sleeping,   0 stopped,   0 zombie
%Cpu(s):  9.2 us,  3.6 sy,  0.0 ni, 86.9 id,  0.1 wa,  0.0 hi,  0.2 si,  0.0 st
GiB Mem :   15.555 total,    1.843 free,    5.044 used,    8.668 buff/cache
GiB Swap:    0.000 total,    0.000 free,    0.000 used.    8.062 avail Mem

  PID USER      PR  NI    VIRT    RES  %CPU %MEM     TIME+ S COMMAND
27648 suresh    20   0 1463.9m 514.3m  31.2   3.2   8:56.20 S chromium

############################

# htop
  1  [||||                                                             3.4%]   5  [||||
  2  [||||                                                             2.7%]   6  [|||||
  3  [||||                                                             3.3%]   7  [|||||
  4  [|||||||                                                          6.0%]   8  [||||||||||||||||||
  Mem[||||||||||||||||||||||||||||||||||||||||||||||||||||||||||7.75G/15.6G]   Tasks: 188; 2 running
  Swp[                                                                0K/0K]   Load average: 1.42 0.96 0.67
                                                                               Uptime: 8 days, 01:21:02

  PID USER      PRI  NI  VIRT   RES   SHR S CPU% MEM%   TIME+  Command
27648 suresh     20   0 1463M  514M  410M S 29.4  3.2  9:07.72 /usr/lib/chromium/chromium --type=renderer --fie

############################
# sudo perf sched record sleep 10
```

Resource: CPU IV

```
# sudo perf sched latency

 -----------------------------------------------------------------------------------------------------------
 Task                  |   Runtime ms  | Switches | Average delay ms | Maximum delay ms | Maximum delay at
 -----------------------------------------------------------------------------------------------------------
 kworker/4:1H:155      |      0.050 ms |        1 | avg:    0.051 ms | max:    0.051 ms | max at: 262585.22959
 khugepaged:67         |      2.600 ms |        1 | avg:    0.042 ms | max:    0.042 ms | max at: 262585.44290
 ksoftirqd/0:7         |      0.116 ms |        9 | avg:    0.042 ms | max:    0.312 ms | max at: 262587.32127
 tr:(20)               |     14.232 ms |       20 | avg:    0.031 ms | max:    0.387 ms | max at: 262580.03030
 kworker/u16:0:31591   |      1.332 ms |       22 | avg:    0.028 ms | max:    0.055 ms | max at: 262580.10994
 sed:(20)              |     55.704 ms |       22 | avg:    0.025 ms | max:    0.211 ms | max at: 262584.03385
 perf:18602            |      1.494 ms |        1 | avg:    0.025 ms | max:    0.025 ms | max at: 262587.63840
 wpa_supplicant:673    |      0.036 ms |        1 | avg:    0.024 ms | max:    0.024 ms | max at: 262583.05633
 WorkerPool/1344:15512 |      0.085 ms |        1 | avg:    0.024 ms | max:    0.024 ms | max at: 262580.73229
 WorkerPool/875:28605  |      2.886 ms |        1 | avg:    0.023 ms | max:    0.023 ms | max at: 262583.96119
 WorkerPool/1553:6841  |      0.266 ms |        1 | avg:    0.021 ms | max:    0.021 ms | max at: 262580.45587
 WorkerPool/1310:536   |      0.135 ms |        1 | avg:    0.021 ms | max:    0.021 ms | max at: 262583.93921
 WorkerPool/696:14642  |      0.169 ms |        2 | avg:    0.021 ms | max:    0.021 ms | max at: 262579.92471
 rcu_preempt:8         |     12.843 ms |      483 | avg:    0.021 ms | max:    0.054 ms | max at: 262581.42652
 WorkerPool/992:9409   |      0.163 ms |        2 | avg:    0.021 ms | max:    0.023 ms | max at: 262580.72839
 mozStorage #3:1398    |     21.238 ms |        2 | avg:    0.020 ms | max:    0.021 ms | max at: 262581.18988
 tmux:(10)             |     12.865 ms |       10 | avg:    0.019 ms | max:    0.128 ms | max at: 262578.03307
 Chrome_DBThread:10992 |      0.446 ms |        6 | avg:    0.019 ms | max:    0.033 ms | max at: 262581.21913
```

[4] http://brendangregg.com/blog/2017-08-08/linux-load-averages.html

Resource: Memory I

- top/htop/atop, smem, vmstat, dmstat etc
- Notes on certain terminology
    - Virtual memory (VSS): Address space used, not an indicative of physical memory usage
    - Resident memory (RSS): Actual physical memory used, including memory shared with other processes. Two kinds of sharing could happen:
        - By forking a process (so both parent and child share same memory), generally with CoW (Copy-on-Write) semantics.
        - By memory mapping same file
    - Unique Set Size (USS): Actual private physical memory used i.e., not including the memory shared with other processes
    - Proportional Set Size (PSS): Private physical memory used + proportion of shared memory with other processes
    - Memory overcommit [5] : Assignment of more memory than physical memory available, assuming not everyone will need all this memory at the same time. For example, Redis needs it:

Resource: Memory II

- Redis bgsave needs it, since bgsave forks a new process out of existing redis process (which is likely using lots of memory), the new forked process also will "appear" to use more memory but in reality it won't need more memory, it just does bgsave and exits.
- So if we disable overcommit, then Redis fork will fail etc.

```
Process A has 50 KiB of unshared memory
Process B has 300 KiB of unshared memory
Both process A and process B have 100 KiB of the same shared memory region

RSS of process A = 50KiB + 100KiB          = 150 KiB
USS of process A                           = 50 KiB
PSS of process A = 50 KiB + (100 KiB / 2)  = 100 KiB

RSS of process B = 300KiB + 100KiB         = 400 KiB
USS of process B                           = 300 KiB
PSS of process B = 300 KiB + (100 KiB / 2) = 350 KiB
```

Some example command invocations:

Resource: Memory III

```
# Total memory

# free -m
             total       used       free     shared    buffers     cached
Mem:          7482       5479       2002          0        154       1573
-/+ buffers/cache:       3751       3731
Swap:            0          0          0

############################

# dstat -m 1
------memory-usage-----
 used  free  buff  cach
8041M 1729M  734M 7682M
8022M 1748M  734M 7662M
8021M 1749M  734M 7662M
8022M 1748M  734M 7661M
8021M 1749M  734M 7661M

############################

# vmstat 1
procs -----------memory---------- ---swap-- -----io---- -system-- ------cpu-----
 r  b   swpd   free   buff  cache   si   so    bi    bo   in   cs us sy id wa st
 1  0      0 1834808 751556 8327776    0    0     8    57   36   53  9  4 87  0  0
 0  0      0 1835048 751556 8327772    0    0     0     0  685 1479  1  1 98  0  0
 0  0      0 1835184 751556 8327516    0    0     0     0  756 1770  1  1 98  0  0
 0  0      0 1835616 751556 8327512    0    0     0     0  669 1710  1  1 98  0  0

############################
```

Resource: Memory IV

```
# top
top - 11:53:53 up 8 days,  1:20,  1 user,  load average: 1.81, 0.96, 0.65
Tasks: 296 total,   1 running, 295 sleeping,   0 stopped,   0 zombie
%Cpu(s):  9.2 us,  3.6 sy,  0.0 ni, 86.9 id,  0.1 wa,  0.0 hi,  0.2 si,  0.0 st
GiB Mem :   15.555 total,    1.843 free,    5.044 used,    8.668 buff/cache
GiB Swap:    0.000 total,    0.000 free,    0.000 used.    8.062 avail Mem

  PID USER      PR  NI    VIRT    RES  %CPU %MEM     TIME+ S COMMAND
27648 suresh    20   0 1463.9m 514.3m  31.2   3.2   8:56.20 S chromium

############################

# htop --sort-key=RES
  1  [                                                            0.0%]   5  [
  2  [                                                            0.0%]   6  [|||||||||||||||||||||||||||||||||
  3  [                                                            0.0%]   7  [
  4  [                                                            0.0%]   8  [
  Mem[|||||||||||||||||||||||||||||||||||||||||||||||||||||7.94G/15.6G]   Tasks: 190; 1 running
  Swp[                                                           0K/0K]   Load average: 0.48 0.53 0.35
                                                                          Uptime: 8 days, 02:00:49

  PID USER      PRI  NI  VIRT   RES   SHR S CPU% MEM%   TIME+  Command
 1223 suresh     20   0 3667M 1334M  267M S  0.0  8.4 36:59.01 /usr/lib/firefox/firefox https://docs.google.com
11054 suresh     20   0 1156M  517M  420M S  0.0  3.2  1:37.12 /usr/lib/chromium/chromium --type=gpu-process --
16413 suresh     20   0 1052M  470M 27264 S  0.0  3.0  7:54.46 ./src/emacs/src/emacs
12621 suresh     20   0 1439M  459M  273M S  0.0  2.9  0:49.31 /usr/lib/chromium/chromium --type=renderer --fie

############################
```

Resource: Memory V

```
# smem -k -t -w
Area                           Used      Cache   Noncache
firmware/hardware                 0          0          0
kernel image                      0          0          0
kernel dynamic memory          1.9G       1.8G      88.2M
userspace memory               3.7G      36.4M       3.7G
free memory                    1.7G       1.7G          0
----------------------------------------------------------
                               7.3G       3.5G       3.8G

# User-wise memory usage
smem -t -k -u
User      Count     Swap      USS      PSS      RSS
rpc           1        0   404.0K   440.0K     2.0M
dbus          1        0   552.0K   586.0K     2.0M
nagios        1        0   692.0K   748.0K     3.1M
rpcuser       1        0   800.0K   862.0K     3.0M
ntp           1        0   764.0K   922.0K     3.9M
smmsp         1        0     1.4M     1.6M     3.4M
nobody        1        0     1.1M     1.9M     7.7M
suresh        2        0   936.0K     2.4M     8.4M
aws           4        0   239.6M   275.3M   394.1M
root         32        0   279.9M   286.1M   353.0M
deploy       19        0     2.9G     3.2G     4.6G
---------------------------------------------------
             64        0     3.4G     3.8G     5.4G
```

Resource: Memory VI

```
# Application-wise memory usage
# smem -k -t
  PID User     Command                         Swap      USS      PSS      RSS
 2971 root     /sbin/mingetty /dev/tty5           0    84.0K   106.0K     1.3M
 2966 root     /sbin/mingetty /dev/tty3           0    88.0K   110.0K     1.4M
 2973 root     /sbin/mingetty /dev/tty6           0    88.0K   110.0K     1.4M
....................
21207 root     Passenger core                     0     6.9M     7.7M    12.8M
18816 root     python /usr/bin/smem -k -t         0     8.1M     8.5M    10.8M
 7154 root     /usr/bin/python /usr/bin/le        0    20.8M    21.2M    24.6M
 2991 aws      opsworks-agent: master 2991        0    18.9M    27.8M    56.5M
 3013 aws      opsworks-agent: statistics         0    52.1M    61.1M    91.2M
 2994 aws      opsworks-agent: keep_alive         0    61.6M    70.6M   100.6M
 3017 aws      opsworks-agent: process_com        0   107.0M   115.8M   145.8M
16816 root     /opt/SumoCollector/jre/bin/        0   228.0M   228.1M   230.3M
21253 deploy   Passenger AppPreloader: /da        0   187.8M   246.8M   474.7M
15144 deploy   Passenger RubyApp: /data/he        0   220.3M   278.4M   501.2M
 7714 deploy   Passenger RubyApp: /data/he        0   281.2M   322.2M   530.4M
23074 deploy   Passenger RubyApp: /data/he        0   328.9M   364.1M   561.7M
20298 deploy   Passenger RubyApp: /data/he        0   531.9M   562.7M   746.9M
20926 deploy   Passenger RubyApp: /data/he        0   658.8M   690.0M   871.8M
16513 deploy   Passenger RubyApp: /data/he        0   753.2M   783.0M   957.8M
-------------------------------------------------------------------------------
   64 11                                          0     3.4G     3.8G     5.4G

# Memory usage by mapping
# smem -k -t -m
Map                                       PIDs   AVGPSS      PSS
/[aio]                                      12        0        0
/data/helpkit/shared/bundler_gems/ruby/2     7        0        0
```

Resource: Memory VII

```
/opt/SumoCollector/19.182-44/lib/aether-     1        0        0
/opt/SumoCollector/19.182-44/lib/aether-     1        0        0
/opt/SumoCollector/19.182-44/lib/akka-ac     1        0        0
....................
/usr/sbin/nginx                             13    58.0K   755.0K
/usr/lib64/perl5/CORE/libperl.so            13    63.0K   823.0K
/usr/lib64/libssl.so.1.0.1k                 32    27.0K   895.0K
/bin/bash                                    2   448.0K   896.0K
/usr/lib64/libnss3.so                       15    80.0K     1.2M
/usr/lib64/libkrb5.so.3.3                   39    33.0K     1.3M
/usr/lib64/libxml2.so.2.9.1                 25    54.0K     1.3M
/usr/lib64/libpython2.7.so.1.0               2   890.0K     1.7M
/usr/local/lib/ruby/gems/2.2.0/bundler/g     3   680.0K     2.0M
/opt/aws/opsworks/local/bin/ruby             4   556.0K     2.2M
[stack]                                     64    35.0K     2.2M
/usr/local/lib/libruby.so.2.2.0              7   371.0K     2.5M
/lib64/libc-2.17.so                         64    41.0K     2.6M
/lib64/libcrypto.so.1.0.1k                  36   101.0K     3.6M
/opt/SumoCollector/jre/lib/amd64/server/     1     6.8M     6.8M
<anonymous>                                 64     4.1M   260.1M
[heap]                                      64    45.7M     2.9G
-----------------------------------------------------------------
448                                       2956    66.5M     3.2G

# See here for possible values: https://www.kernel.org/doc/Documentation/vm/overcommit-accounting
# sysctl vm.overcommit_memory
vm.overcommit_memory = 0

# Out of memory errors
dmesg -T | grep OOM
```

[5] https://www.etalabs.net/overcommit.html

Resource: Disk I

- Linux file system architecture

```
+----------------------------+
|      Physical Machine      |
| +------------------------+ |
| |   +----------------+   | |
| |   |  Application   |   | |
| |   +-------+--------+   | |
| | UserSpace |            | |
| +-----------+------------+ |
| | Kernel    |            | |
| |   +-------+--------+   | |
| |   | Syscall Layer  |   | |
| |   +----------------+   | |
| |   |  VFS Layer     |   | |
| |   +----+--+---+----+   | |
| |   |ext4|  |xfs|        | |
| |   +----+--+---+-+      | |
| |   | Page Cache  |      | |
| |   +-------------+--+   | |
| |   |  Block Layer   |   | |
| |   +----------------+   | |
| |   | Device Driver  |   | |
| |   +----------------+   | |
| +------------------------+ |
|   +------------------+     |
|   |      Disk        |     |
|   +------------------+     |
+----------------------------+
```

Resource: Disk II

- Typical/Simplified workflow of how disk writes are done (assuming without O_DIRECT or O_SYNC)
    - Application makes write(somedata) syscall
    - Data is transferred to kernel page cache (page cache == Unused RAM is used to cache the data read/written)
    - write() call returns
    - After 'sometime', data is transferred to disk
    - Notes
        - The write() call can block at times. For example, when the page cache is full (vm.dirty_ratio in below example)
        - The data is written asynchronously. How often is based on some parameters and/or cache/buffer status.
        - Either writeback time is reached (vm.dirty_writeback_centisecs or vm.dirtytime_expire_seconds) or the page cache is full.
- Linux file system architecture when virtualized (xen)

Resource: Disk III

```
+-------------------------------------------------------------------------------------------------------------------------------+
|                                                      Physical Machine                                                         |
| Dom0                             DomU (Driver Domain)             DomU (Guest)                     DomU (Guest)               |
| +------------------------+       +------------------------+       +------------------------+      +------------------------+  |
| |  +---------------+     |       |                        |       |   +----------------+   |      |   +----------------+   |  |
| |  |Xen Control S/W|     |       |                        |       |   |  Application   |   |      |   |  Application   |   |  |
| |  +---------------+     |       |                        |       |   +-------+--------+   |      |   +-------+--------+   |  |
| | UserSpace              |       | UserSpace              |       | UserSpace |            |      | UserSpace |            |  |
| +------------------------+       +------------------------+       +-----------+------------+      +-----------+------------+  |
| | Kernel                 |       | Kernel                 |       | Kernel    |            |      | Kernel    |            |  |
| |                        |       |                        |       |   +-------+--------+   |      |   +-------+--------+   |  |
| |                        |       |                        |       |   | Syscall Layer  |   |      |   | Syscall Layer  |   |  |
| |                        |       |                        |       |   +----------------+   |      |   +----------------+   |  |
| |                        |       |                        | ....  |   |  VFS Layer     |   | .... |   |  VFS Layer     |   |  |
| |                        |       |                        |       |   +----+--+---+----+   |      |   +----+--+---+----+   |  |
| |                        |       |                        |       |   |ext4|  |xfs|        |      |   |ext4|  |xfs|        |  |
| |                        |       |                        |       |   +----+--+---+-+      |      |   +----+--+---+-+      |  |
| |                        |       |                        |       |   | Page Cache  |      |      |   | Page Cache  |      |  |
| |  +----------------+    |       |   +----------------+   |       |   +-------------+--+   |      |   +-------------+--+   |  |
| |  | Backend Driver |    |       |   | Backend Driver +---+---+   |   |  Block Layer   |   |      |   |  Block Layer   |   |  |
| |  +----------------+    |       |   +----------------+   |   |   |   +----------------+   |      |   +----------------+   |  |
| |  | Native Driver  |    |       |   | Native Driver  |   |   +---+---+ Frontend Driver|   |      |   | Frontend Driver|   |  |
| |  +----+-----------+    |       |   +-------+--------+   |   |   |   +----------------+   |      |   +-------+--------+   |  |
| |       |                |       |           |            |   |   |                        |      |           |            |  |
| +-------+----------------+       +-----------+------------+   |   +------------------------+      +-----------+------------+  |
|         |                                    |                |                                               |               |
|         +------------------------------------+                +-----------------------------------------------+               |
|         |                                                                                                                     |
|         |                 +-----------------------------------------------------------------+                                  |
|         |                 |Hypervisor                                                       |                                  |
|         |                 |   +---------------+  +----------+  +-------------+ +---------+  |                                  |
|         |                 |   |vCPU 1...vCPU N|  |VirtualMMU|  |Event Channel| |Scheduler|  |                                  |
|         |                 |   +---------------+  +----------+  +-------------+ +---------+  |                                  |
|         |                 +-----------------------------------------------------------------+                                  |
|         |                                                                                                                     |
|  +------+-------------------------------------------------------------------------------------------------------------------+ |
|  | Hardware                                                                                                                 | |
|  |   +-----+   +-----+   +------+   +------+   +------+    +-----+   +-----+                                                | |
|  |   |CPU 1|...|CPU N|   |Memory|   |Disk 1|...|Disk N|    |NIC 1|...|NIC N|  ....                                          | |
|  |   +-----+   +-----+   +------+   +----- +   +------+    +-----+   +-----+                                                | |
|  +--------------------------------------------------------------------------------------------------------------------------+ |
+-------------------------------------------------------------------------------------------------------------------------------+
```

Resource: Disk IV

- Impact of virtualization [6]
  - Virtualization adds extra layer of redirection, increasing latency and bottleneck
  - Additionally, network based disks (AWS EBS, EFS etc) bring-in variation in performance that we cannot control and/or measure at times.
    - AWS specific: Use EBS optimized instances, supposed to have separate/dedicated NIC for EBS traffic (Ref?)
  - Performance may vary significantly due to multi-tenancy / noisy-neighbor
  - For example, a high IOWait time may / may not have any relation with our IOPS (== noisy-neighbor saturating local disk controller or network card in case of network file system)
- Life of a byte in Disk IO
  We will write one byte into a file and then follow that byte as it flows through various subsystems

Resource: Disk V

```
# Life of a byte

# Install kernel debug info

yum-config-manager --enable "amzn-main-debuginfo" --enable "amzn-updates-debuginfo"
yum -y install kernel-debuginfo kernel-devel

# Setup the device
mkfs.ext4 /dev/xvdc
mount /dev/xvdc /tmp/test

# Check the block size
tune2fs -l /dev/xvdc | grep -i 'block size'
Block size:               4096

# Create a file with just one byte
echo -n "n" > /tmp/test/foo

############################

# Create probe points: vfs layer, block layer and then from xen-blkfront driver ("drivers/block/xen-blkfront.c")
perf probe --add='vfs_*' --add='blkif_*' --add='blkfront_*' --add='blkback_*' --add='xlvbd_*'

# Run the file write operation with probes enabled
# We will open the above file with O_SYNC flag, and then just update one byte in it
perf trace -T --event 'block:*' --event='probe:vfs_*' --event='probe:blkif_*' --event='probe:blkfront_*' \
 --event='probe:blkback_*' --event='probe:xlvbd_*' --event='ext4:*' \
  ruby -e 'f=open("/tmp/test/foo", File::RDWR + File::SYNC); f.write("y"); f.close()'

78224651.658 ( 0.031 ms): ruby/5215 brk(                                                                  ) = 0x11e8000
78224651.719 ( 0.033 ms): ruby/5215 mmap(len: 4096, prot: READ|WRITE, flags: PRIVATE|ANONYMOUS, fd: -1    ) = 0x7fd9b3080000
78224651.774 ( 0.020 ms): ruby/5215 access(filename: 0xb2e7e490, mode: R                                  ) = -1 ENOENT No such file
78224651.814 ( 0.020 ms): ruby/5215 open(filename: 0xb2e7cd95, flags: CLOEXEC                             ) ...
...........
...........
78224789.337 ( 0.031 ms): ruby/5215 open(filename: 0x15e2810, flags: CLOEXEC|RDWR|SYNC|0x101000           ) ...
78224789.368 (         ): probe:vfs_open:(ffffffff811f87b0))
78224789.337 ( 0.062 ms): ruby/5215  ... [continued]: open()) = 7
78224789.426 ( 0.025 ms): ruby/5215 fcntl(fd: 7, cmd: GETFD, arg: 7                                        ) = 1
78224789.481 ( 0.026 ms): ruby/5215 fstat(fd: 7, statbuf: 0x7ffcccbc42f0                                   ) ...
78224789.508 (         ): probe:vfs_fstat:(ffffffff811fefc0))
78224789.532 (         ): probe:vfs_getattr:(ffffffff811fef90))
```

Resource: Disk VI

```
78224789.558 (         ): probe:vfs_getattr_nosec:(ffffffff811fee60))
78224789.481 ( 0.108 ms): ruby/5215  ... [continued]: fstat()) = 0
78224789.614 ( 0.025 ms): ruby/5215 ioctl(fd: 7, cmd: TCGETS, arg: 0x7ffcccbc4340                         ) = -1 ENOTTY Inappropriate
78224789.670 ( 0.029 ms): ruby/5215 write(fd: 7, buf: 0x15e3fe0, count: 1                                  ) ...
78224789.699 (         ): probe:vfs_write:(ffffffff811fa070))
78224789.731 (         ): ext4:ext4_journal_start:dev 202,32 blocks, 2 rsv_blocks, 0 caller ext4_dirty_inode)
78224789.762 (         ): ext4:ext4_mark_inode_dirty:dev 202,32 ino 12 caller ext4_dirty_inode)
78224789.791 (         ): block:block_touch_buffer:202,32 sector=1057 size=4096)
78224789.827 (         ): ext4:ext4_da_write_begin:dev 202,32 ino 12 pos 0 len 1 flags 0)
78224789.855 (         ): ext4:ext4_journal_start:dev 202,32 blocks, 1 rsv_blocks, 0 caller ext4_da_write_begin)
78224789.885 (         ): ext4:ext4_da_write_end:dev 202,32 ino 12 pos 0 len 1 copied 1)
78224789.914 (         ): block:block_dirty_buffer:202,32 sector=34304 size=4096)
78224789.967 (         ): probe:vfs_fsync_range:(ffffffff8122c1f0))
78224789.990 (         ): ext4:ext4_sync_file_enter:dev 202,32 ino 12 parent 2 datasync 0 )
78224790.020 (         ): ext4:ext4_writepages:dev 202,32 ino 12 nr_to_write 9223372036854775807 pages_skipped 0 range_start 0 range_
78224790.048 (         ): ext4:ext4_journal_start:dev 202,32 blocks, 8 rsv_blocks, 0 caller ext4_writepages)
78224790.109 (         ): ext4:ext4_da_write_pages:dev 202,32 ino 12 first_page 0 nr_to_write 9223372036854775807 sync_mode 1)
78224790.142 (         ): block:block_bio_queue:202,32 WS 274432 + 8 [ruby])
78224790.170 (         ): block:block_getrq:202,32 WS 274432 + 8 [ruby])
78224790.199 (         ): block:block_plug:[ruby])
78224790.221 (         ): block:block_rq_insert:202,32 WS 0 () 274432 + 8 [ruby])
78224790.235 (         ): block:block_unplug:[ruby] 1)
78224790.237 (         ): block:block_rq_issue:202,32 WS 0 () 274432 + 8 [ruby])
78224790.238 (         ): probe:blkif_queue_request:(ffffffff8143f970))
78224790.239 (         ): probe:blkif_ring_get_request:(ffffffff8143d810))
78224790.240 (         ): probe:blkif_setup_rw_req_grant:(ffffffff814435d0))
78224790.315 (         ): ext4:ext4_writepages_result:dev 202,32 ino 12 ret 0 pages_written 1 pages_skipped 0 sync_mode 1 writeback_i
...........
...........
78224792.127 (         ): ext4:ext4_sync_file_exit:dev 202,32 ino 12 ret 0)
78224789.670 ( 2.503 ms): ruby/5215  ... [continued]: write()) = 1
78224792.203 ( 0.030 ms): ruby/5215 close(fd: 7                                                           ) = 0
...........
...........
78224803.693 ( 0.000 ms): ruby/5215 exit_group(                                                           )

############################

# Looking above operations from PoV of block layer

btrace /dev/xvdc
202,32   0        1    43.474871183  5657  Q   R 270344 + 8 [ruby]
```

Resource: Disk VII

```
202,32   0        2    43.474872807  5657  G   R 270344 + 8 [ruby]
202,32   0        3    43.474873303  5657  I   R 270344 + 8 [ruby]
202,32   0        4    43.474873914  5657  D   R 270344 + 8 [ruby]
202,32   2        2    43.475358283     0  C   R 270344 + 8 [0]
202,32   2        3    43.475382340  5657  Q  WS 270344 + 8 [ruby]
202,32   2        4    43.475383344  5657  G  WS 270344 + 8 [ruby]
202,32   2        5    43.475383618  5657  P   N [ruby]
202,32   2        6    43.475384342  5657  I  WS 270344 + 8 [ruby]
202,32   2        7    43.475384678  5657  U   N [ruby] 1
202,32   2        8    43.475384971  5657  D  WS 270344 + 8 [ruby]
202,32   2        9    43.475922196     0  C  WS 270344 + 8 [0]
202,32   2       10    43.475931892  5657  Q WSM 8456 + 8 [ruby]
202,32   2       11    43.475932544  5657  G WSM 8456 + 8 [ruby]
202,32   2       12    43.475932855  5657  I WSM 8456 + 8 [ruby]
202,32   2       13    43.475933142  5657  D WSM 8456 + 8 [ruby]
202,32   2       14    43.476441001     0  C WSM 8456 + 8 [0]

So looks like 8 sectors, starting from 270344 were written.
Why?

# Lets get the file details

stat /tmp/test/foo
  File: ‘/tmp/test/foo’
  Size: 1            Blocks: 8          IO Block: 4096   regular file
Device: ca20h/51744d    Inode: 12          Links: 1
.........
.........

# So it is 8 sectors (because block size is 4096)

# Get the file’s sector details
hdparm --fibmap /tmp/test/foo

/tmp/test/foo:
 filesystem blocksize 4096, begins at LBA 0; assuming 512 byte sectors.
 byte_offset  begin_LBA    end_LBA    sectors
           0     270344     270351          8

# So it is indeed sector 270344 that is where the file is stored
# Check if we have the single byte ’y’ we wrote stored there.
dd if=/dev/xvdc bs=512 skip=270344 count=1 status=none | hexdump -C
```

Resource: Disk VIII

```
00000000  79 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00  |y...............|
00000010  00 00 00 00 00 00 00 00  00 00 00 00 00 00 00 00  |................|
*
00000200
```

- Disk utilization at system level / per device

```
Check system load, high load avg might indicate disk utilization/saturation as well
# uptime
 04:48:56 up 173 days, 21:51,  2 users,  load average: 85.01, 84.57, 83.81

############################

Check disk utilization/saturation by device
# iostat -xz 1
Linux 4.4.44-39.55.amzn1.x86_64 (cluster-2-data-108)    08/11/2017      _x86_64_        (8 CPU)

avg-cpu:  %user   %nice %system %iowait  %steal   %idle
           1.56    0.02    0.18    0.66    0.01   97.57

Device:         rrqm/s   wrqm/s     r/s     w/s   rsec/s   wsec/s avgrq-sz avgqu-sz   await  svctm  %util
xvda              0.00     2.68    0.07    2.31     2.59    42.58    18.91     0.00    1.65   0.95   0.23
xvdi              1.47     1.07   13.86   16.79  1121.99  2295.66   111.52     0.19    6.18   0.97   2.97
xvdj              1.47     1.31   13.86   18.66  1121.92  2311.07   105.56     0.02    6.64   0.96   3.11
md0               0.00     0.00   26.90   53.28  2243.91  4606.73    85.44     0.00    0.00   0.00   0.00
dm-0              0.00     0.00   20.97   36.57  2243.91  4606.73   119.07     0.08    3.67   1.01   5.79

############################
```

Resource: Disk IX

```
# dstat 1
----total-cpu-usage---- -dsk/total- -net/total- ---paging-- ---system--
usr sys idl wai hiq siq| read  writ| recv  send|  in   out | int   csw
  2   0  98   1   0   0|1125k 2346k|   0     0 |   0     0 | 980  1253
  2   0   0  98   0   0|   0   160k|6642B 4859B|   0     0 | 881  1173
  0   0   0 100   0   0|   0     0 |  66B  158B|   0     0 | 245   448
  0   0   0 100   0   0|   0     0 | 426B 6394B|   0     0 | 295   510
  0   0   0 100   0   0|   0     0 | 164B  216B|   0     0 | 252   457
  0   0   0 100   0   0|   0     0 | 737B  632B|   0     0 | 698   906
  0   0   0 100   0   0|   0     0 | 164B  216B|   0     0 | 594   888
  0   0   0 100   0   0|   0     0 |2355B 2426B|   0     0 | 411   633
  0   0   0 100   0   0|   0     0 | 328B 7872B|   0     0 | 303   511
  0   0   0 100   0   0|   0     0 |  66B  126B|   0     0 | 710   939
  1   1   0  98   0   0|   0   208k|2643B 1566B|   0     0 |2063  2309
  1   0   0  99   0   0|   0     0 |2136B 2313B|   0     0 | 810  1145
  0   0   0 100   0   0|   0     0 | 295B 1428B|   0     0 | 801   955
  0   0   0 100   0   0|   0     0 | 240B 7806B|   0     0 | 292   508
  0   0   0 100   0   0|   0     0 |2266B 2297B|   0     0 | 402   638
  0   0   0 100   0   0|   0    88k| 639B  534B|   0     0 | 560   830
  0   0   0 100   0   0|   0   280k|  66B  134B|   0     0 | 814  1066

############################

# vmstat 1
procs -----------memory---------- ---swap-- -----io---- --system-- -----cpu-----
 r  b   swpd   free   buff  cache   si   so    bi    bo   in   cs us sy id wa st
 0 21      0 452900  99716 27156232    0    0   141   291    0    0  2  0 98  1  0
 0 21      0 453132  99716 27156232    0    0     0     0  452  743  0  0  0 100  0
 0 21      0 453132  99720 27156232    0    0     0    48  267  480  0  0  0 100  0
```

Resource: Disk X

```
0 21      0 452460  99720 27156232    0    0     0     0 1261 1403  0  0  0 99  0
0 21      0 452512  99720 27156232    0    0     0     0  715  944  0  0  0 100  0
0 21      0 452512  99720 27156232    0    0     0     4  320  535  0  0  0 100  0
0 21      0 452636  99720 27156232    0    0     0     0  504  834  0  0  0 100  0
0 21      0 452636  99724 27156228    0    0     0    20  251  457  0  0  0 100  0
0 21      0 452636  99724 27156232    0    0     0     0  321  484  0  0  0 100  0
0 21      0 452636  99724 27156232    0    0     0     0  303  529  0  0  0 100  0
0 21      0 452760  99724 27156232    0    0     0     0  705  941  0  0  0 100  0
0 21      0 452124  99724 27156232    0    0     0     0 1192 1463  0  0  0 99  0
0 21      0 452124  99724 27156232    0    0     0    60  265  472  0  0  0 100  0
0 21      0 452008  99724 27156232    0    0     0    16  920 1051  0  1  0 99  0
0 21      0 452008  99724 27156232    0    0     0     0  348  603  0  0  0 100  0
0 21      0 452008  99724 27156232    0    0     0     0  242  442  0  0  0 100  0
```

- Disk utilization by process

Resource: Disk XI

```
# iotop -o
Total DISK READ: 0.00 B/s | Total DISK WRITE: 7.94 K/s
  TID  PRIO  USER     DISK READ  DISK WRITE  SWAPIN     IO>    COMMAND
 7720 be/4 deploy      0.00 B/s    3.97 K/s  0.00 %  0.00 % Passenger RubyApp: /data/helpkit/current/public (pr
25654 be/4 root        0.00 B/s   47.65 K/s  0.00 %  0.00 % Passenger core
25655 be/4 root        0.00 B/s    0.00 B/s  0.00 %  0.00 % Passenger core
21647 be/4 root        0.00 B/s    3.97 K/s  0.00 %  0.00 % Passenger core
21229 be/4 deploy      0.00 B/s    3.97 K/s  0.00 %  0.00 % nginx: worker process
20932 be/4 deploy      0.00 B/s   35.74 K/s  0.00 %  0.00 % Passenger RubyApp: /data/helpkit/current/public (pr
15151 be/4 deploy      0.00 B/s    7.94 K/s  0.00 %  0.00 % Passenger RubyApp: /data/helpkit/current/public (pr
25656 be/4 root        0.00 B/s   55.59 K/s  0.00 %  0.00 % Passenger core
25657 be/4 root        0.00 B/s   51.62 K/s  0.00 %  0.00 % Passenger core
23081 be/4 deploy      0.00 B/s   51.62 K/s  0.00 %  0.00 % Passenger RubyApp: /data/helpkit/current/public (pr
16853 be/4 root        0.00 B/s    7.94 K/s  0.00 %  0.00 % java -XX:+UseParallelGC -server -Xms64m -Xmx128m -D

############################

# pidstat -d
Linux 4.4.51-40.67.amzn1.x86_64 (rails-app-4)   08/09/2017      _x86_64_        (4 CPU)

07:54:05 AM       PID   kB_rd/s   kB_wr/s kB_ccwr/s  Command
07:54:05 AM         1     36.06     39.42      6.94  init
07:54:05 AM        31      0.00      0.00      0.00  xenwatch
07:54:05 AM      1571      0.00      3.95      0.00  jbd2/xvda1-8
07:54:05 AM      1614      0.00      0.00      0.00  udevd
07:54:05 AM      2439      0.00      0.00      0.00  dhclient
07:54:05 AM      2548      0.00      0.00      0.00  dhclient
07:54:05 AM      2595      0.00      0.20      0.00  auditd
..............
07:53:36 AM     21204      0.00      0.00      0.00  PassengerAgent
```

Resource: Disk XII

```
07:53:36 AM     21207      0.00     29.44     28.89  PassengerAgent
07:53:36 AM     21214      0.00      0.00      0.00  PassengerAgent
07:53:36 AM     21223      0.00      0.00      0.00  nginx
07:53:36 AM     21230      0.00      0.42      0.28  nginx
07:53:36 AM     21231      0.00      0.41      0.26  nginx
07:53:36 AM     21234      0.00      0.44      0.29  nginx
07:53:36 AM     21235      0.00      0.43      0.28  nginx
07:53:36 AM     21236      0.00      0.38      0.23  nginx
07:53:36 AM     21253      0.50     37.51      2.33  ruby
07:53:36 AM     23074      0.00      0.03      0.00  ruby
07:53:36 AM     25853      0.00      0.00      0.00  pidstat
```

- Find processes that are in uninterruptible state (most likely due to disk IO)

Resource: Disk XIII

Processes that are in uninterruptible state

```
# ps axl | awk ’$10 ~ /[D]/’
0     0   654   653  20   0 118448  1548 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
1     0   677     2  20   0      0     0 -      D    ?         59:19 [kswapd0]
0     0  1024  1021  20   0 118448  1588 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0  1255  1251  20   0 118448  1428 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
1     0  1319     2  20   0      0     0 -      D    ?          0:12 [kworker/1:0]
1     0  1671     2   0 -20      0     0 -      D<   ?          0:23 [kworker/2:2H]
0     0  1933  1929  20   0 118448  1464 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0  3458  3453  20   0 118448  1536 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0  4058  4057  20   0 118448  1424 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0  4159  4156  20   0 118448  1536 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0  5739  5737  20   0 118448  1548 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0  5775  5773  20   0 118448  1516 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0  5786  5781  20   0 118448  1460 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0  6293  6292  20   0 118448  1520 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0  6501  6497  20   0 118448  1452 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
1     0  6686     2  20   0      0     0 -      D    ?         47:52 [xfsaild/dm-0]
0     0  8147  8142  20   0 118448  1424 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0  8589  8586  20   0 118448  1548 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0  8784  8779  20   0 118448  1536 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0  9463  9460  20   0 118448  1588 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 10988 10986  20   0 118448  1460 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 11606 11603  20   0 118448  1584 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 11695 11694  20   0 118448  1520 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 12189 12188  20   0 118448  1584 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 12288 12283  20   0 118448  1584 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 13248 13246  20   0 118448  1552 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 13932 13929  20   0 118448  1588 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 14153 14151  20   0 118448  1424 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
1     0 15115     2  20   0      0     0 -      D    ?          0:09 [kworker/2:2]
0     0 15724 15721  20   0 118448  1532 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 16157 16156  20   0 118448  1584 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 16374 16371  20   0 118448  1456 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 17458 17453  20   0 118448  1512 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 18530 18529  20   0 118448  1524 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 18591 18588  20   0 118448  1552 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 18628 18625  20   0 118448  1460 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 18637 18632  20   0 118448  1372 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 19324 19319  20   0 118448  1460 -      Ds   ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 20835 20833  30  10 120572  2316 -      DN   ?          0:00 /usr/sbin/logrotate /etc/logrotate.conf
```

Resource: Disk XIV

```
0     0 20877 20876  20   0 118448  1588 -      Ds    ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 21452 21450  20   0 118448  1528 -      Ds    ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 21624 21619  20   0 118448  1380 -      Ds    ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 23113 23110  20   0 118448  1520 -      Ds    ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 23676 23673  20   0 118448  1588 -      Ds    ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 23881 23876  20   0 118448  1516 -      Ds    ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 25271 25270  20   0 118448  1368 -      Ds    ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 25477 25474  20   0 118448  1588 -      Ds    ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 25828 25826  20   0 118448  1512 -      Ds    ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 26043 26041  20   0 118448  1548 -      Ds    ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 26167 26162  20   0 118448  1380 -      Ds    ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 26774 26773  20   0 118448  1512 -      Ds    ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 28326 28325  20   0 118448  1520 -      Ds    ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 28981 28978  20   0 118448  1480 -      Ds    ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 28982 28977  20   0 118448  1584 -      Ds    ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 30585 30584  20   0 118448  1516 -      Ds    ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 30912 30909  20   0 118448  2196 -      Ds    ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 31209 31206  20   0 118448  1512 -      Ds    ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 31456 31451  20   0 118448  1532 -      Ds    ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit
0     0 31816 31812  20   0 118448  1588 -      Ds    ?          0:00 /usr/sbin/logrotate /etc/logrotate.d/goaudit

############################

What are they waiting on?

# ps axl | awk ’$10 ~ /[D]/’ | sudo awk ’{ print "====="$13"==="$3"====="; system("cat /proc/"$3"/stack")}’
=====/usr/sbin/logrotate===654=====
[<ffffffff812daa54>] call_rwsem_down_read_failed+0x14/0x30
[<ffffffffa03b303f>] xfs_ilock+0xff/0x130 [xfs]
[<ffffffffa03b30a0>] xfs_ilock_data_map_shared+0x30/0x40 [xfs]
[<ffffffffa03a7060>] xfs_dir_open+0x30/0x60 [xfs]
[<ffffffff811d2e23>] do_dentry_open+0x223/0x300
[<ffffffff811d40e5>] vfs_open+0x55/0x80
[<ffffffff811e17e0>] path_openat+0x1b0/0x12a0
[<ffffffff811e467e>] do_filp_open+0x7e/0xd0
[<ffffffff811d4468>] do_sys_open+0x128/0x210
[<ffffffff811d4584>] SyS_openat+0x14/0x20
[<ffffffff814efcae>] entry_SYSCALL_64_fastpath+0x12/0x71
[<ffffffffffffffff>] 0xffffffffffffffff
=====[kswapd0]===677=====
[<ffffffff81083f5f>] flush_work+0xef/0x170
[<ffffffffa03c5659>] xlog_cil_force_lsn+0x79/0x1e0 [xfs]
```

Resource: Disk XV

```
[<ffffffffa03c3cd1>] _xfs_log_force_lsn+0x71/0x310 [xfs]
[<ffffffffa03c3f9e>] xfs_log_force_lsn+0x2e/0xa0 [xfs]
[<ffffffffa03b2b8d>] __xfs_iunpin_wait+0x8d/0x140 [xfs]
[<ffffffffa03b6329>] xfs_iunpin_wait+0x19/0x20 [xfs]
[<ffffffffa03ab722>] xfs_reclaim_inode+0x122/0x340 [xfs]
[<ffffffffa03abb54>] xfs_reclaim_inodes_ag+0x214/0x330 [xfs]
[<ffffffffa03ac773>] xfs_reclaim_inodes_nr+0x33/0x40 [xfs]
[<ffffffffa03bb099>] xfs_fs_free_cached_objects+0x19/0x20 [xfs]
[<ffffffff811d85c1>] super_cache_scan+0x181/0x190
[<ffffffff81172a56>] shrink_slab.part.41+0x206/0x3f0
[<ffffffff81176809>] shrink_zone+0x2a9/0x2c0
[<ffffffff81177794>] kswapd+0x4b4/0x960
[<ffffffff8108a7a9>] kthread+0xc9/0xe0
[<ffffffff814f000f>] ret_from_fork+0x3f/0x70
[<ffffffffffffffff>] 0xffffffffffffffff
=====/usr/sbin/logrotate===1024=====
[<ffffffff812daa54>] call_rwsem_down_read_failed+0x14/0x30
[<ffffffffa03b303f>] xfs_ilock+0xff/0x130 [xfs]
[<ffffffffa03b30a0>] xfs_ilock_data_map_shared+0x30/0x40 [xfs]
[<ffffffffa03a7060>] xfs_dir_open+0x30/0x60 [xfs]
[<ffffffff811d2e23>] do_dentry_open+0x223/0x300
[<ffffffff811d40e5>] vfs_open+0x55/0x80
[<ffffffff811e17e0>] path_openat+0x1b0/0x12a0
[<ffffffff811e467e>] do_filp_open+0x7e/0xd0
[<ffffffff811d4468>] do_sys_open+0x128/0x210
[<ffffffff811d4584>] SyS_openat+0x14/0x20
[<ffffffff814efcae>] entry_SYSCALL_64_fastpath+0x12/0x71
[<ffffffffffffffff>] 0xffffffffffffffff
=====/usr/sbin/logrotate===1255=====
[<ffffffff812daa54>] call_rwsem_down_read_failed+0x14/0x30
[<ffffffffa03b303f>] xfs_ilock+0xff/0x130 [xfs]
[<ffffffffa03b30a0>] xfs_ilock_data_map_shared+0x30/0x40 [xfs]
[<ffffffffa03a7060>] xfs_dir_open+0x30/0x60 [xfs]
[<ffffffff811d2e23>] do_dentry_open+0x223/0x300
[<ffffffff811d40e5>] vfs_open+0x55/0x80
[<ffffffff811e17e0>] path_openat+0x1b0/0x12a0
[<ffffffff811e467e>] do_filp_open+0x7e/0xd0
[<ffffffff811d4468>] do_sys_open+0x128/0x210
[<ffffffff811d4584>] SyS_openat+0x14/0x20
[<ffffffff814efcae>] entry_SYSCALL_64_fastpath+0x12/0x71
[<ffffffffffffffff>] 0xffffffffffffffff
=====[kworker/1:0]===1319=====
```

Resource: Disk XVI

```
[<ffffffff810b2b51>] down+0x41/0x50
[<ffffffffa03a2afc>] xfs_buf_lock+0x3c/0xf0 [xfs]
[<ffffffffa03a2d12>] _xfs_buf_find+0x162/0x340 [xfs]
[<ffffffffa03a2f1a>] xfs_buf_get_map+0x2a/0x280 [xfs]
[<ffffffffa03a3bdd>] xfs_buf_read_map+0x2d/0x180 [xfs]
[<ffffffffa03cf664>] xfs_trans_read_buf_map+0xf4/0x310 [xfs]
[<ffffffffa037c329>] xfs_btree_read_buf_block.constprop.28+0x69/0xa0 [xfs]
[<ffffffffa037c3d1>] xfs_btree_lookup_get_block+0x71/0xe0 [xfs]
[<ffffffffa0380c37>] xfs_btree_lookup+0xb7/0x560 [xfs]
[<ffffffffa0367091>] xfs_free_ag_extent+0x61/0x760 [xfs]
[<ffffffffa03687ea>] xfs_free_extent+0xda/0x110 [xfs]
[<ffffffffa03cff16>] xfs_trans_free_extent+0x26/0x60 [xfs]
[<ffffffffa039f74f>] xfs_bmap_finish+0xff/0x120 [xfs]
[<ffffffffa03b5453>] xfs_itruncate_extents+0x113/0x240 [xfs]
[<ffffffffa03a0384>] xfs_free_eofblocks+0x1b4/0x210 [xfs]
[<ffffffffa03accf5>] xfs_inode_free_eofblocks+0x95/0x160 [xfs]
[<ffffffffa03ab2ce>] xfs_inode_ag_walk.isra.10+0x1ee/0x310 [xfs]
[<ffffffffa03ac531>] xfs_inode_ag_iterator_tag+0x71/0xa0 [xfs]
[<ffffffffa03ac7fd>] xfs_icache_free_eofblocks+0x2d/0x40 [xfs]
[<ffffffffa03ac82b>] xfs_eofblocks_worker+0x1b/0x30 [xfs]
[<ffffffff81084ba0>] process_one_work+0x150/0x3f0
[<ffffffff8108531a>] worker_thread+0x11a/0x470
[<ffffffff8108a7a9>] kthread+0xc9/0xe0
[<ffffffff814f000f>] ret_from_fork+0x3f/0x70
[<ffffffffffffffff>] 0xffffffffffffffff
=====[kworker/2:2H]===1671=====
[<ffffffff81083f5f>] flush_work+0xef/0x170
[<ffffffffa03c5659>] xlog_cil_force_lsn+0x79/0x1e0 [xfs]
[<ffffffffa03c3986>] _xfs_log_force+0x76/0x270 [xfs]
[<ffffffffa03c3ba6>] xfs_log_force+0x26/0x90 [xfs]
[<ffffffffa03c3c34>] xfs_log_worker+0x24/0x50 [xfs]
[<ffffffff81084ba0>] process_one_work+0x150/0x3f0
[<ffffffff8108531a>] worker_thread+0x11a/0x470
[<ffffffff8108a7a9>] kthread+0xc9/0xe0
[<ffffffff814f000f>] ret_from_fork+0x3f/0x70
[<ffffffffffffffff>] 0xffffffffffffffff
...........
...........
...........
=====/usr/sbin/logrotate===31816=====
[<ffffffff812daa54>] call_rwsem_down_read_failed+0x14/0x30
[<ffffffffa03b303f>] xfs_ilock+0xff/0x130 [xfs]
```

Resource: Disk XVII

```
[<ffffffffa03b30a0>] xfs_ilock_data_map_shared+0x30/0x40 [xfs]
[<ffffffffa03a7060>] xfs_dir_open+0x30/0x60 [xfs]
[<ffffffff811d2e23>] do_dentry_open+0x223/0x300
[<ffffffff811d40e5>] vfs_open+0x55/0x80
[<ffffffff811e17e0>] path_openat+0x1b0/0x12a0
[<ffffffff811e467e>] do_filp_open+0x7e/0xd0
[<ffffffff811d4468>] do_sys_open+0x128/0x210
[<ffffffff811d4584>] SyS_openat+0x14/0x20
[<ffffffff814efcae>] entry_SYSCALL_64_fastpath+0x12/0x71
[<ffffffffffffffff>] 0xffffffffffffffff
```

- Find disk activity at block IO layer

```
Find the activity at block layer level
NOTE: Use btt tool for extended analysis: http://www.cse.unsw.edu.au/~aaronc/iosched/doc/btt.html
# btrace == blktrace /dev/xvda -o - | blkparse -s -i -
# btrace /dev/xvda
202,0    3        1     0.000000000 11720  A   W 6257952 + 8 <- (202,1) 6253856
202,0    3        2     0.000000904 11720  Q   W 6257952 + 8 [java]
202,0    3        3     0.000004811 11720  G   W 6257952 + 8 [java]
202,0    3        4     0.000005466 11720  P   N [java]
.............
.............
202,0    3       55     1.337693195  1571  A  WS 4584696 + 8 <- (202,1) 4580600
202,0    3       56     1.337693283  1571  Q  WS 4584696 + 8 [jbd2/xvda1-8]
202,0    3       57     1.337693403  1571  M  WS 4584696 + 8 [jbd2/xvda1-8]
202,0    3       58     1.337693622  1571  A  WS 4584704 + 8 <- (202,1) 4580608
202,0    3       59     1.337693710  1571  Q  WS 4584704 + 8 [jbd2/xvda1-8]
.............
.............
202,0    3      126     2.001851710 11720  Q   W 6256664 + 8 [java]
202,0    3      127     2.001855937 11720  G   W 6256664 + 8 [java]
202,0    3      128     2.001856577 11720  P   N [java]
202,0    3      129     2.001859526 11720  I   W 6256664 + 8 [java]
202,0    3      130     2.001860479 11720  U   N [java] 1
202,0    3      131     2.001861857 11720  D   W 6256664 + 8 [java]
```

Resource: Disk XVIII

```
202,0    3      132     2.002029389 11720  A   W 6261568 + 8 <- (202,1) 6257472
202,0    3      133     2.002029792 11720  Q   W 6261568 + 8 [java]
202,0    3      134     2.002031071 11720  G   W 6261568 + 8 [java]
202,0    3      135     2.002031370 11720  P   N [java]
202,0    3      136     2.002032645 11720  I   W 6261568 + 8 [java]
202,0    3      137     2.002033044 11720  U   N [java] 1
202,0    3      138     2.002033469 11720  D   W 6261568 + 8 [java]
202,0    3      139     2.002453955     0  C   W 6256664 + 8 [0]
202,0    3      140     2.002516093     0  C   W 6261568 + 8 [0]
202,0    3      141     3.002859806 11720  A   W 6255424 + 8 <- (202,1) 6251328
202,0    3      142     3.002860373 11720  Q   W 6255424 + 8 [java]
202,0    3      143     3.002862233 11720  G   W 6255424 + 8 [java]
202,0    3      144     3.002862542 11720  P   N [java]
202,0    3      145     3.002864546 11720  I   W 6255424 + 8 [java]
202,0    3      146     3.002864960 11720  U   N [java] 1
.............
.............
202,0    3      147     3.002865474 11720  D   W 6255424 + 8 [java]
202,0    3      268    10.009390888 11720  D   W 6257960 + 8 [java]
202,0    3      269    10.009769306     0  C   W 6257952 + 8 [0]
202,0    3      270    10.009931136     0  C   W 6257960 + 8 [0]
^C
 java (11720)
 Reads Queued:           0,        0KiB  Writes Queued:         22,       88KiB
 Read Dispatches:        0,        0KiB  Write Dispatches:      22,       88KiB
 Reads Requeued:         0               Writes Requeued:        0
 Reads Completed:        0,        0KiB  Writes Completed:       0,        0KiB
 Read Merges:            0,        0KiB  Write Merges:           0,        0KiB
 IO unplugs:            22               Timer unplugs:          0
 Allocation wait:        0               Allocation wait:        0
 Dispatch wait:          0               Dispatch wait:          0
 Completion wait:        0               Completion wait:        0
jbd2/xvda1-8 (1571)
 Reads Queued:           0,        0KiB  Writes Queued:         46,      184KiB
 Read Dispatches:        0,        0KiB  Write Dispatches:       4,      184KiB
 Reads Requeued:         0               Writes Requeued:        0
 Reads Completed:        0,        0KiB  Writes Completed:       0,        0KiB
 Read Merges:            0,        0KiB  Write Merges:          42,      168KiB
 IO unplugs:             2               Timer unplugs:          0
 Allocation wait:        0               Allocation wait:        0
 Dispatch wait:          0               Dispatch wait:          0
 Completion wait:        0               Completion wait:        0
```

Resource: Disk XIX

```
swapper/3 (0)
 Reads Queued:           0,        0KiB  Writes Queued:            0,        0KiB
 Read Dispatches:        0,        0KiB  Write Dispatches:         0,        0KiB
 Reads Requeued:         0               Writes Requeued:          0
 Reads Completed:        0,        0KiB  Writes Completed:        24,      264KiB
 Read Merges:            0,        0KiB  Write Merges:             0,        0KiB
 IO unplugs:             0               Timer unplugs:            0
 Allocation wait:        0               Allocation wait:          0
 Dispatch wait:          0               Dispatch wait:            0
 Completion wait:        0               Completion wait:          0
utils.rb:110 (6680)
 Reads Queued:           0,        0KiB  Writes Queued:            0,        0KiB
 Read Dispatches:        0,        0KiB  Write Dispatches:         0,        0KiB
 Reads Requeued:         0               Writes Requeued:          0
 Reads Completed:        0,        0KiB  Writes Completed:         2,        8KiB
 Read Merges:            0,        0KiB  Write Merges:             0,        0KiB
 IO unplugs:             0               Timer unplugs:            0
 Allocation wait:        0               Allocation wait:          0
 Dispatch wait:          0               Dispatch wait:            0
 Completion wait:        0               Completion wait:          0

.............
.............

Throughput (R/W): 0KiB/s / 27KiB/s
Events (202,0): 330 entries
Skips: 0 forward (0 -   0.0%)
```

- Using blktrace to trace/observe the activity at block layer

Resource: Disk XX

```
$ btrace /dev/xvdz                               || $ echo 3 > /proc/sys/vm/drop_caches
202,6400  1 1   0.0000 15982  Q   R 0 + 32 [dd]  || $ dd if=/dev/xvdz bs=512 of=/dev/null count=1
202,6400  1 2   0.0000 15982  G   R 0 + 32 [dd]  || 1+0 records in
202,6400  1 3   0.0000 15982  P   N [dd]         || 1+0 records out
202,6400  1 4   0.0000 15982  I   R 0 + 32 [dd]  || 512 bytes (512 B) copied, 0.000728468 s, 703 kB/s
202,6400  1 5   0.0000 15982  U   N [dd] 1       ||
202,6400  1 6   0.0000 15982  D   R 0 + 32 [dd]  || $ dd if=/dev/xvdz bs=512 of=/dev/null count=2
202,6400  3 1   0.0005      0  C   R 0 + 32 [0]   || 2+0 records in
                                                 || 2+0 records out
                                                 || 1024 bytes (1.0 kB) copied, 8.07e-05 s, 12.7 MB/s
                                                 ||
                                                 || $ dd if=/dev/xvdz bs=512 of=/dev/null count=8
                                                 || 8+0 records in
                                                 || 8+0 records out
                                                 || 4096 bytes (4.1 kB) copied, 0.000111268 s, 36.8 MB/s
                                                 ||
                                                 ||
                                                 ||
202,6400  0 1  32.8711 16110  Q   R 32 + 64 [dd] || $ dd if=/dev/xvdz bs=512 of=/dev/null count=9
202,6400  0 2  32.8711 16110  G   R 32 + 64 [dd  || 9+0 records in
202,6400  0 3  32.8711 16110  P   N [dd]         || 9+0 records out
202,6400  0 4  32.8711 16110  I   R 32 + 64 [dd  || 4608 bytes (4.6 kB) copied, 0.000149306 s, 30.9 MB/s
202,6400  0 5  32.8711 16110  U   N [dd] 1       ||
202,6400  0 6  32.8711 16110  D   R 32 + 64 [dd  ||
202,6400  3 2  32.8719      0  C   R 32 + 64 [0]  ||
                                                 ||
                                                 ||
202,6400  2 1 147.7486 17283  Q   R 0 + 1 [dd]   || $ dd if=/dev/xvdz bs=512 of=/dev/null count=1 iflag=direct
202,6400  2 2 147.7486 17283  G   R 0 + 1 [dd]   || 1+0 records in
202,6400  2 3 147.7486 17283  P   N [dd]         || 1+0 records out
202,6400  2 4 147.7486 17283  I   R 0 + 1 [dd]   || 512 bytes (512 B) copied, 0.000728468 s, 703 kB/s
202,6400  2 5 147.7486 17283  U   N [dd] 1       ||
202,6400  2 6 147.7486 17283  D   R 0 + 1 [dd]   ||
202,6400  3 3 147.7490  9973  C   R 0 + 1 [0]
```

Resource: Disk XXI

- Example trace out of a "bad" disk

```
# perf trace --event 'block:*' dd if=/dev/xvdj of=/dev/null bs=512 count=1 iflag=direct
     0.175 ( 0.016 ms): dd/28637 brk(                                                                  ) = 0x10be000
     0.221 ( 0.018 ms): dd/28637 mmap(len: 4096, prot: READ|WRITE, flags: PRIVATE|ANONYMOUS, fd: -1    ) = 0x7f1f56c20000
     0.252 ( 0.015 ms): dd/28637 access(filename: 0x56a1f140, mode: R                                  ) = -1 ENOENT No such file or
     0.282 ( 0.016 ms): dd/28637 open(filename: 0x56a1da38, flags: CLOEXEC                             ) = 3
     0.307 ( 0.012 ms): dd/28637 fstat(fd: 3, statbuf: 0x7ffc2c5e33b0                                  ) = 0
     ...............
     ...............
     1.638 ( 0.025 ms): dd/28637 open(filename: 0x2c5e5739, flags: CREAT|TRUNC|WRONLY, mode: 438       ) = 3
     1.700 ( 0.039 ms): dd/28637 dup2(oldfd: 3, newfd: 1                                               ) = 1
     1.728 ( 0.013 ms): dd/28637 close(fd: 3                                                           ) = 0
     1.759 ( 0.017 ms): dd/28637 clock_gettime(which_clock: MONOTONIC, tp: 0x7ffc2c5e3b40              ) = 0
     1.809 ( 0.033 ms): dd/28637 read(buf: 0x10c0000, count: 512                                       ) ...
     1.809 (         ): block:block_bio_queue:202,144 R 0 + 1 [dd])
     1.836 (         ): block:block_getrq:202,144 R 0 + 1 [dd])
     1.858 (         ): block:block_plug:[dd])
     1.875 (         ): block:block_rq_insert:202,144 R 0 () 0 + 1 [dd])
     1.887 (         ): block:block_unplug:[dd] 1)
^C
```

- Bad disk(s) can have cascading effect on unrelated disk activity as well

Resource: Disk XXII

```
# Ran 'yum install' on a system that had bad disk (but rootfs disk was fine)
# Yum install got stuck after about 80% work done
# Analyzing the where it is stuck showed the below stack:
#   When it tried to allocate a page out of page cache,
#   it ran out of free pages (or reached water mark), so it tried to reclaim
#   pages, which led to the trying to sync pages belonging to 'bad' disk (xfs
#   in this case), causing it to be stuck

cat /proc/`pidof yum`/stack
[<ffffffff81083f5f>] flush_work+0xef/0x170
[<ffffffffa03c5659>] xlog_cil_force_lsn+0x79/0x1e0 [xfs]
[<ffffffffa03c3cd1>] _xfs_log_force_lsn+0x71/0x310 [xfs]
[<ffffffffa03c3f9e>] xfs_log_force_lsn+0x2e/0xa0 [xfs]
[<ffffffffa03b2b8d>] __xfs_iunpin_wait+0x8d/0x140 [xfs]
[<ffffffffa03b6329>] xfs_iunpin_wait+0x19/0x20 [xfs]
[<ffffffffa03ab722>] xfs_reclaim_inode+0x122/0x340 [xfs]
[<ffffffffa03abb54>] xfs_reclaim_inodes_ag+0x214/0x330 [xfs]
[<ffffffffa03ac773>] xfs_reclaim_inodes_nr+0x33/0x40 [xfs]
[<ffffffffa03bb099>] xfs_fs_free_cached_objects+0x19/0x20 [xfs]
[<ffffffff811d85c1>] super_cache_scan+0x181/0x190
[<ffffffff81172a56>] shrink_slab.part.41+0x206/0x3f0
[<ffffffff81176809>] shrink_zone+0x2a9/0x2c0
[<ffffffff81176ba5>] do_try_to_free_pages+0x175/0x440
[<ffffffff81176f25>] try_to_free_pages+0xb5/0x170
[<ffffffff8116abaa>] __alloc_pages_nodemask+0x53a/0xa60
[<ffffffff811aef58>] alloc_pages_current+0x88/0x120
[<ffffffff81162294>] __page_cache_alloc+0xb4/0xc0
[<ffffffff81162c76>] pagecache_get_page+0x56/0x1e0
[<ffffffff81162e26>] grab_cache_page_write_begin+0x26/0x40
[<ffffffffa0120e01>] ext4_da_write_begin+0xa1/0x330 [ext4]
[<ffffffff81161e50>] generic_perform_write+0xc0/0x1a0
[<ffffffff81163f48>] __generic_file_write_iter+0x188/0x1e0
[<ffffffffa0115b76>] ext4_file_write_iter+0xf6/0x360 [ext4]
[<ffffffff811d4c5a>] __vfs_write+0xaa/0xe0
[<ffffffff811d5282>] vfs_write+0xa2/0x1a0
[<ffffffff811d5f86>] SyS_write+0x46/0xa0
[<ffffffff814efcae>] entry_SYSCALL_64_fastpath+0x12/0x71
[<ffffffffffffffff>] 0xffffffffffffffff
```

Resource: Disk XXIII

- Disk space usage

```
# df -h
Filesystem       Size  Used Avail Use% Mounted on
devtmpfs         3.7G   64K  3.7G   1% /dev
tmpfs            3.7G     0  3.7G   0% /dev/shm
/dev/xvda1       7.8G  3.4G  4.3G  44% /
/dev/xvdh         99G  5.1G   89G   6% /data
```

- Disk related errors

```
# dmesg -T | grep "blocked for more than"
INFO: task xfsaild/dm-0:6686 blocked for more than 120 seconds.

# demsg -T | grep "I/O error"
[351410.715652] EXT4-fs warning (device xvdh): htree_dirblock_to_tree:958: inode #262145: lblock 0: comm ls: er
[397736.767853] blk_update_request: I/O error, dev xvdh, sector 73992
[397736.770649] EXT4-fs warning (device xvdh): htree_dirblock_to_tree:958: inode #2: lblock 0: comm ls: error -
[399503.066719] blk_update_request: I/O error, dev xvdh, sector 73992
```

[6] http://dtrace.org/blogs/brendan/2013/01/11/virtualization-performance-zones-kvm-xen/

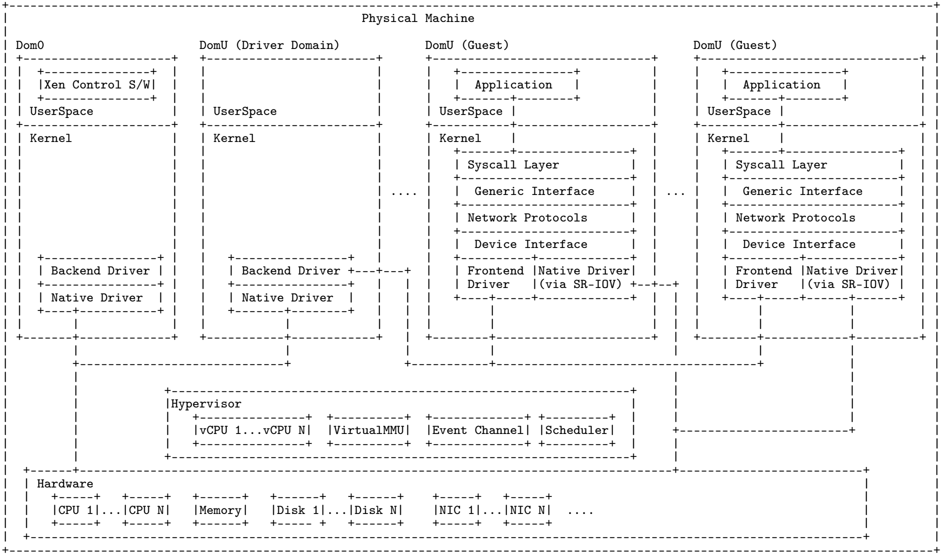
Resource: Network I

- Linux network stack architecture

```
+------------------------------+
|       Physical Machine       |
|                              |
| +--------------------------+ |
| |   +----------------+     | |
| |   |  Application   |     | |
| |   +-------+--------+     | |
| | UserSpace |              | |
| +-----------+--------------+ |
| | Kernel    |              | |
| |   +-------+------------+ | |
| |   | Syscall Layer      | | |
| |   +--------------------+ | |
| |   |  Generic Interface   | |
| |   +--------------------+ | |
| |   | Network Protocols  | | |
| |   +--------------------+ | |
| |   |  Device Interface  | | |
| |   +--------------------+ | |
| |   | Device Driver (NIC)| | |
| |   +--------------------+ | |
| +--------------------------+ |
|     +------------------+     |
|     |        NIC       |     |
|     +------------------+     |
+------------------------------+
```

- Linux when stack architecture when virtualized (xen)

Resource: Network II



```
+-----------------------------------------------------------------------------------------------------------------------------------+
|                                                  Physical Machine                                                                 |
|                                                                                                                                   |
| Dom0                      DomU (Driver Domain)            DomU (Guest)                          DomU (Guest)                      |
| +---------------------+   +------------------------+      +-------------------------------+     +-------------------------------+ |
| |  +---------------+  |   |                        |      |   +----------------+          |     |   +----------------+          | |
| |  |Xen Control S/W|  |   |                        |      |   |  Application   |          |     |   |  Application   |          | |
| |  +---------------+  |   |                        |      |   +-------+--------+          |     |   +-------+--------+          | |
| | UserSpace           |   | UserSpace              |      | UserSpace |                   |     | UserSpace |                   | |
| +---------------------+   +------------------------+      +-----------+-------------------+     +-----------+-------------------+ |
| | Kernel              |   | Kernel                 |      | Kernel    |                   |     | Kernel    |                   | |
| |                     |   |                        |      |   +-------+----------------+  |     |   +-------+----------------+  | |
| |                     |   |                        |      |   | Syscall Layer          |  |     |   | Syscall Layer          |  | |
| |                     |   |                        |      |   +------------------------+  |     |   +------------------------+  | |
| |                     |   |                        | .... |   |  Generic Interface     |  | ... |   |  Generic Interface     |  | |
| |                     |   |                        |      |   +------------------------+  |     |   +------------------------+  | |
| |                     |   |                        |      |   | Network Protocols      |  |     |   | Network Protocols      |  | |
| |                     |   |                        |      |   +------------------------+  |     |   +------------------------+  | |
| |                     |   |                        |      |   |  Device Interface      |  |     |   |  Device Interface      |  | |
| |  +----------------+ |   |   +----------------+   |      |   +----------+-------------+  |     |   +----------+-------------+  | |
| |  | Backend Driver | |   |   | Backend Driver +---+---+  |   | Frontend |Native Driver|  |     |   | Frontend |Native Driver|  | |
| |  +----------------+ |   |   +----------------+   |   |  |   | Driver   |(via SR-IOV) +--+--+  |   | Driver   |(via SR-IOV) |  | |
| |  | Native Driver  | |   |   | Native Driver  |   |   |  |   +----+-----+-------------+  |  |  |   +----+-----+------+------+  | |
| |  +----+-----------+ |   |   +-------+--------+   |   |  |        |                      |  |  |        |            |         | |
| |       |             |   |           |            |   |  |        |                      |  |  |        |            |         | |
| +-------+-------------+   +-----------+------------+   |  +--------+----------------------+  |  +--------+------------+---------+ |
|         |                             |                |           |                         |           |            |           |
|         +-----------------------------+                +-----------+-------------------------------------+            |           |
|         |                                                                                    |                        |           |
|         |            +-----------------------------------------------------------------+     |                        |           |
|         |            |Hypervisor                                                       |     |                        |           |
|         |            |   +---------------+  +----------+  +-------------+ +---------+  |     |                        |           |
|         |            |   |vCPU 1...vCPU N|  |VirtualMMU|  |Event Channel| |Scheduler|  |     +------------------------+           |
|         |            |   +---------------+  +----------+  +-------------+ +---------+  |     |                                    |
|         |            +-----------------------------------------------------------------+     |                                    |
|  +------+------------------------------------------------------------------------------------+--------------------------+         |
|  | Hardware                                                                                                             |         |
|  |   +-----+   +-----+   +------+   +------+   +------+    +-----+   +-----+                                            |         |
|  |   |CPU 1|...|CPU N|   |Memory|   |Disk 1|...|Disk N|    |NIC 1|...|NIC N|  ....                                      |         |
|  |   +-----+   +-----+   +------+   +----- +   +------+    +-----+   +-----+                                            |         |
|  +----------------------------------------------------------------------------------------------------------------------+         |
+-----------------------------------------------------------------------------------------------------------------------------------+
```

Resource: Network III

- Impact of virtualization
    - Virtualization adds extra layer of redirection, increasing latency and bottleneck
    - Performance may vary significantly due to multi-tenancy / noisy-neighbor
        - For example, sudden high latency/throughput drop may / may not have any relation with our network traffic itself (== noisy-neighbor saturating local NIC controller or switch)
    - Use SR-IOV enabled network device if available. This will allow the guest OS to directly talk to the hardware, without going through the Driver Domain
        - AWS specific: Enhanced network support

Resource: Network IV

- Life of a byte in network stack
  We will send a simple HTTP GET request and trace as it goes through various subsystems
  Youc can find in below links more detailed walk through of various network layers: https://blog.packagecloud.io/eng/2016/06/22/monitoring-tuning-linux-networking-stack-receiving-data/ https://blog.packagecloud.io/eng/2017/02/06/monitoring-tuning-linux-networking-stack-sending-data/

```
Install kernel debug info
# yum-config-manager --enable "amzn-main-debuginfo" --enable "amzn-updates-debuginfo"
# yum -y install kernel-debuginfo kernel-devel

Find the kernel module responsible for our ethernet card
# lspci -k
....................
....................
00:03.0 Ethernet controller: Intel Corporation 82599 Ethernet Controller Virtual Function (rev 01)
        Kernel driver in use: ixgbevf
....................
....................

############################

Get all the probeable functions in our driver
# perf probe -m ixgbevf  -F ’ixgbevf_*’
ixgbevf_addr_list_itr
.....
```

Resource: Network V

```
.....
ixgbevf_xmit_frame

Create probe point for all these functions
# perf probe -m ixgbevf --add=’ixgbevf_*’
Too many( > 128) probe point found.
Added new events:
  probe:ixgbevf_addr_list_itr (on ixgbevf_* in ixgbevf)
  probe:ixgbevf_set_rx_mode (on ixgbevf_* in ixgbevf)
  probe:ixgbevf_vlan_rx_kill_vid (on ixgbevf_* in ixgbevf)
  probe:ixgbevf_change_mtu (on ixgbevf_* in ixgbevf)
  probe:ixgbevf_set_mac (on ixgbevf_* in ixgbevf)
  probe:ixgbevf_negotiate_api (on ixgbevf_* in ixgbevf)
  probe:ixgbevf_free_q_vector (on ixgbevf_* in ixgbevf)
  probe:ixgbevf_free_q_vectors (on ixgbevf_* in ixgbevf)
  probe:ixgbevf_free_irq (on ixgbevf_* in ixgbevf)
  probe:ixgbevf_update_itr (on ixgbevf_* in ixgbevf)
  .......
  .......
  probe:ixgbevf_init_module (on ixgbevf_* in ixgbevf)
  probe:ixgbevf_exit_module (on ixgbevf_* in ixgbevf)

You can now use it in all perf tools, such as:

        perf record -e probe:ixgbevf_exit_module -aR sleep 1

##################################################################

##################################################################################

Now trace the packets, simple ping command tracing

# perf trace --event ’net:*’ --event ’probe:ixgbevf_*’ ping -c1 8.8.8.8 >/dev/null
     0.196 ( 0.027 ms): ping/3910 brk(                                                                   ) = 0x5564c6b32000
     ............
     ............
     5.204 ( 0.032 ms): ping/3910 sendmsg(fd: 3<socket:[87842167]>, msg: 0x5564c6a68160                 ) ...
     5.204 (         ): net:net_dev_queue:dev=eth0 skbaddr=0xffff8800eaf35800 len=98)
     5.229 (         ): net:net_dev_start_xmit:dev=eth0 queue_mapping=0 skbaddr=0xffff8800eaf35800 vlan_tagged=0 vlan_proto=0x0000 vl
     5.253 (         ): probe:ixgbevf_xmit_frame:(ffffffffa025a670))
     5.276 (         ): net:net_dev_xmit:dev=eth0 skbaddr=0xffff8800eaf35800 len=98 rc=0)
     5.299 ( 0.126 ms): ping/3910  ... [continued]: sendmsg()) = 64
```

Resource: Network VI

```
     5.346 ( 0.024 ms): ping/3910 setitimer(which: REAL, value: 0x7fff20936fe0                           ) = 0
    18.001 (12.632 ms): ping/3910 recvmsg(fd: 3<socket:[87842167]>, msg: 0x7fff20937000                  ) ...
    18.001 (         ): probe:ixgbevf_msix_clean_rings:(ffffffffa0259b20))
    18.043 (         ): probe:ixgbevf_poll:(ffffffffa025c080))
    ............
    ............
    18.285 (12.915 ms): ping/3910  ... [continued]: recvmsg()) = 84
    18.294 (         ): probe:ixgbevf_msix_clean_rings:(ffffffffa0259b20))
    18.296 (         ): probe:ixgbevf_poll:(ffffffffa025c080))
    18.297 (         ): probe:ixgbevf_clean_rx_irq:(ffffffffa0259b70))
    18.298 (         ): probe:ixgbevf_update_itr:(ffffffffa02594e0))
    18.302 (         ): net:net_dev_queue:dev=eth0 skbaddr=0xffff8801e3a0fce8 len=166)
    18.304 (         ): net:net_dev_start_xmit:dev=eth0 queue_mapping=0 skbaddr=0xffff8801e3a0fce8 vlan_tagged=0 vlan_proto=0x0000 vl
    18.305 (         ): probe:ixgbevf_xmit_frame:(ffffffffa025a670))
    18.306 (         ): net:net_dev_xmit:dev=eth0 skbaddr=0xffff8801e3a0fce8 len=166 rc=0)
    ............
    ............
    18.546 (         ): probe:ixgbevf_poll:(ffffffffa025c080))
    18.571 (         ): probe:ixgbevf_clean_rx_irq:(ffffffffa0259b70))
    18.597 (         ): net:napi_gro_receive_entry:dev=eth0 napi_id=0x1 queue_mapping=0 skbaddr=0xffff8801e6bd4d00 vlan_tagged=0 vlan
    18.621 (         ): net:netif_receive_skb:dev=eth0 skbaddr=0xffff8801e6bd4d00 len=52)
    18.649 (         ): net:napi_gro_receive_entry:dev=eth0 napi_id=0x1 queue_mapping=0 skbaddr=0xffff8801e6bd4d00 vlan_tagged=0 vlan
    18.675 (         ): net:netif_receive_skb:dev=eth0 skbaddr=0xffff8801e6bd4d00 len=52)
    18.701 (         ): probe:ixgbevf_alloc_rx_buffers:(ffffffffa0259750))
    ............
    ............
    18.810 (         ): probe:ixgbevf_poll:(ffffffffa025c080))
    18.835 (         ): probe:ixgbevf_clean_rx_irq:(ffffffffa0259b70))
    18.860 (         ): net:napi_gro_receive_entry:dev=eth0 napi_id=0x1 queue_mapping=0 skbaddr=0xffff8801e6bd4d00 vlan_tagged=0 vlan
    18.885 (         ): net:netif_receive_skb:dev=eth0 skbaddr=0xffff8801e6bd4d00 len=52)
    ............
    ............
    19.141 ( 0.024 ms): ping/3910 write(fd: 1</dev/null>, buf: 0x7fe157a07000, count: 99                 ) = 99
    19.186 ( 0.023 ms): ping/3910 write(fd: 1</dev/null>, buf: 0x7fe157a07000, count: 1                  ) = 1
    19.243 ( 0.023 ms): ping/3910 write(fd: 1</dev/null>, buf: 0x7fe157a07000, count: 145                ) = 145
    19.265 ( 0.000 ms): ping/3910 exit_group(

##################################################################################
# Now we will trace a HTTP request

# We will start and establish the TCP connection, wait for keyboard input
# and then send the HTTP GET request.
```

Resource: Network VII

```
# We will only start tracing after connection establishment, so we can just
# focus on GET request alone

              TERMINAL 1                                                TERMINAL 2

(read  -n 1 -p "Press any key to continue "; \        ||  # tshark -f "not port 22"
  echo -e -n ’GET / HTTP/1.1\r\n’ \                    ||
  echo -e -n ’Host: support.freshdesk.com\r\n\r\n’) \  ||  Running as user "root" and group "root". This could be dangerous.
 | socat -t 10 - TCP4:support.freshdesk.com:80         ||  Capturing on eth0
Press any key to continue                              ||
                                                       ||  0.000000000 172.23.3.135 -> 172.23.0.2   DNS 81 Standard query 0xc5dc  A s
                                                       ||  0.000234617   172.23.0.2 -> 172.23.3.135 DNS 113 Standard query response 0
                                                       ||  0.000341919 172.23.3.135 -> 52.206.84.26 TCP 74 43074 > http [SYN] Seq=0 W
                                                       ||  0.001194278 52.206.84.26 -> 172.23.3.135 TCP 74 http > 43074 [SYN, ACK] Se
                                                       ||  0.001208949 172.23.3.135 -> 52.206.84.26 TCP 66 43074 > http [ACK] Seq=1 A
                                                       ||
                                                       ||
                                                       ||
HTTP/1.1 302 Found                                     ||  9.942741717 172.23.3.135 -> 52.206.84.26 HTTP 113 GET / HTTP/1.1
Cache-Control: no-cache                                ||  9.943058252 172.23.3.135 -> 52.206.84.26 TCP 66 43074 > http [FIN, ACK] Se
Content-Type: text/html; charset=utf-8                 ||  9.943200815 52.206.84.26 -> 172.23.3.135 TCP 66 http > 43074 [ACK] Seq=1 A
Date: Tue, 05 Sep 2017 08:37:21 GMT                    ||  9.964304451 52.206.84.26 -> 172.23.3.135 HTTP 600 HTTP/1.1 302 Found  (tex
Location: https://support.freshdesk.com/               ||  9.964317286 172.23.3.135 -> 52.206.84.26 TCP 66 43074 > http [ACK] Seq=49
Set-Cookie: _x_w=1; path=/                             ||  9.964319806 52.206.84.26 -> 172.23.3.135 TCP 66 http > 43074 [FIN, ACK] Se
Status: 302 Found                                      ||  9.964322372 172.23.3.135 -> 52.206.84.26 TCP 66 43074 > http [ACK] Seq=49
X-Frame-Options: SAMEORIGIN                            ||
X-Rack-Cache: miss                                     ||
X-Request-Id: 794f2c16bc159a2dd339b0c33cc4394d         ||
X-Runtime: 0.018734                                    ||
X-UA-Compatible: IE=Edge,chrome=1                      ||
X-XSS-Protection: 1; mode=block                        ||
Content-Length: 96                                     ||
Connection: Close                                      ||
                                                       ||
<html><body>You are being                              ||
<a href="https://support.freshdesk.com/">redirected    ||
</a>.</body></html>                                    ||


TERMINAL 3

perf trace -T  --event ’net:*’ --event ’probe:vfs*’ --event ’probe:ixgbevf_*’ -p ‘pidof socat‘
```

Resource: Network VIII

```
..............
..............
     0.000 ( 0.000 ms):  ... [continued]: select()) = 1
686549238.395 ( 0.033 ms): read(buf: 0x1fc6040, count: 8192                                      ) ...
686549238.427 (         ): probe:vfs_read:(ffffffff811f9f40))
686549238.395 ( 0.061 ms):  ... [continued]: read()) = 47
686549238.483 ( 0.028 ms): write(fd: 3<socket:[354463]>, buf: 0x1fc6040, count: 47               ) ...
686549238.512 (         ): probe:vfs_write:(ffffffff811fa070))
686549238.543 (         ): net:net_dev_queue:dev=eth0 skbaddr=0xffff8801c87550e8 len=113)
686549238.578 (         ): net:net_dev_start_xmit:dev=eth0 queue_mapping=0 skbaddr=0xffff8801c87550e8 vlan_tagged=0 vlan_proto=0x0000
686549238.605 (         ): probe:ixgbevf_xmit_frame:(ffffffffa01d9820))
686549238.633 (         ): net:net_dev_xmit:dev=eth0 skbaddr=0xffff8801c87550e8 len=113 rc=0)
686549238.483 ( 0.179 ms):  ... [continued]: write()) = 47
686549238.692 ( 0.028 ms): select(n: 4, inp: 0x7ffe7ef65f60, outp: 0x7ffe7ef65fe0, exp: 0x7ffe7ef66060) = 2
686549238.747 ( 0.026 ms): read(buf: 0x1fc6040, count: 8192                                      ) ...
686549238.772 (         ): probe:vfs_read:(ffffffff811f9f40))
686549238.747 ( 0.052 ms):  ... [continued]: read()) = 0
686549238.828 ( 0.030 ms): shutdown(fd: 3<socket:[354463]>, how: 1                               ) ...
686549238.859 (         ): net:net_dev_queue:dev=eth0 skbaddr=0xffff8801c8756ae8 len=66)
686549238.894 (         ): net:net_dev_start_xmit:dev=eth0 queue_mapping=0 skbaddr=0xffff8801c8756ae8 vlan_tagged=0 vlan_proto=0x0000
686549238.920 (         ): probe:ixgbevf_xmit_frame:(ffffffffa01d9820))
686549238.946 (         ): net:net_dev_xmit:dev=eth0 skbaddr=0xffff8801c8756ae8 len=66 rc=0)
686549238.828 ( 0.147 ms):  ... [continued]: shutdown()) = 0
686549239.003 (21.167 ms): select(n: 4, inp: 0x7ffe7ef65f60, outp: 0x7ffe7ef65fe0, exp: 0x7ffe7ef66060, tvp: 0x7ffe7ef66170) = 1
686549260.213 ( 0.025 ms): read(fd: 3<socket:[354463]>, buf: 0x1fc6040, count: 8192              ) ...
686549260.239 (         ): probe:vfs_read:(ffffffff811f9f40))
686549260.213 ( 0.056 ms):  ... [continued]: read()) = 534
686549260.297 ( 0.027 ms): write(fd: 1</dev/pts/4>, buf: 0x1fc6040, count: 534                   ) ...
686549260.324 (         ): probe:vfs_write:(ffffffff811fa070))
686549260.297 ( 0.065 ms):  ... [continued]: write()) = 534
686549260.389 ( 0.027 ms): shutdown(fd: 3<socket:[354463]>, how: 1                               ) = -1 ENOTCONN Transport endpoint i
686549260.447 ( 0.029 ms): select(n: 4, inp: 0x7ffe7ef65f60, outp: 0x7ffe7ef65fe0, exp: 0x7ffe7ef66060, tvp: 0x7ffe7ef66170) = 2
686549260.503 ( 0.027 ms): read(fd: 3<socket:[354463]>, buf: 0x1fc6040, count: 8192              ) ...
686549260.530 (         ): probe:vfs_read:(ffffffff811f9f40))
686549260.503 ( 0.053 ms):  ... [continued]: read()) = 0
686549260.584 ( 0.030 ms): shutdown(fd: 3<socket:[354463]>, how: 1                               ) = -1 ENOTCONN Transport endpoint i
686549260.649 ( 0.029 ms): ioctl(fd: 1</dev/pts/4>, cmd: TCSETS, arg: 0x7ffe7ef65f70             ) = 0
686549260.704 ( 0.026 ms): ioctl(fd: 1</dev/pts/4>, cmd: TCGETS, arg: 0x7ffe7ef65f70             ) = 0
686549260.758 ( 0.026 ms): shutdown(fd: 3<socket:[354463]>, how: 2                               ) = -1 ENOTCONN Transport endpoint i
686549260.881 ( 0.000 ms): exit_group(                                                           )
```

Resource: Network IX

```
# Delete the probes
perf probe -m ixgbevf --del='ixgbevf_*'
perf probe --del='vfs_*'
```

- Utilization at system level

```
# Utilization

# nethogs
NetHogs version 0.8.5

    PID USER     PROGRAM                                            DEV    SENT      RECEIVED
   1281 deploy   ..ssenger RubyApp: /data/helpkit/current/public (prod eth0   27.519   279.612 KB/sec
  25051 deploy   ..ssenger RubyApp: /data/helpkit/current/public (prod eth0   33.842   279.514 KB/sec
   3044 deploy   ..ssenger RubyApp: /data/helpkit/current/public (prod eth0   12.543    86.839 KB/sec
  20292 deploy   ..ssenger RubyApp: /data/helpkit/current/public (prod eth0   36.001    86.133 KB/sec
  21235 deploy   nginx: worker process                              eth0   98.778    22.356 KB/sec
   7154 root     /usr/bin/python                                    eth0  122.532     3.029 KB/sec
   8073 deploy   ..ssenger RubyApp: /data/helpkit/current/public (prod eth0    0.801     2.462 KB/sec
   9090 deploy   ..ssenger RubyApp: /data/helpkit/current/public (prod eth0    0.454     1.387 KB/sec
   9410 root     tail                                               eth0    0.392     0.381 KB/sec
  16816 root     /opt/SumoCollector/jre/bin/java                    eth0    1.734     0.380 KB/sec
  12620 suresh   sshd: suresh@pts/0                                 eth0    0.178     0.052 KB/sec
      ? root     10.2.16.117:45404-54.231.141.84:443                        0.000     0.000 KB/sec
      ? root     10.2.16.117:53200-10.2.204.10:9101                         0.000     0.000 KB/sec
      ? root     unknown TCP                                                0.000     0.000 KB/sec

  TOTAL                                                                   334.773   762.145 KB/sec
```

Resource: Network X

```
############################

# iftop

                   1.91Mb              3.81Mb             5.72Mb              7.63Mb        9.54Mb
---------------------------------------------------------------------------------------------------
rails-app-4.localdomain             => ip-10-2-86-132.eu-west-1.compute.i   460Kb   310Kb   310Kb
                                    <=                                     8.64Mb  5.09Mb  5.09Mb
rails-app-4.localdomain             => ip-10-2-86-150.eu-west-1.compute.i   429Kb   247Kb   247Kb
                                    <=                                     5.77Mb  3.25Mb  3.25Mb
rails-app-4.localdomain             => ec2-34-253-108-119.eu-west-1.compu  2.06Mb  1.48Mb  1.48Mb
                                    <=                                     28.2Kb  25.2Kb  25.2Kb
rails-app-4.localdomain             => ip-10-2-20-57.eu-west-1.compute.in  96.1Kb  77.0Kb  77.0Kb
                                    <=                                      745Kb   807Kb   807Kb
rails-app-4.localdomain             => ip-10-2-10-9.eu-west-1.compute.int   729Kb   519Kb   519Kb
                                    <=                                      385Kb   255Kb   255Kb
rails-app-4.localdomain             => ip-10-2-21-95.eu-west-1.compute.in  86.6Kb  79.1Kb  79.1Kb
                                    <=                                      559Kb   682Kb   682Kb
rails-app-4.localdomain             => collector-3.newrelic.com                0b    415Kb   415Kb
                                    <=                                         0b   9.27Kb  9.27Kb
rails-app-4.localdomain             => ip-10-2-10-48.eu-west-1.compute.in   156Kb   207Kb   207Kb
                                    <=                                      121Kb   114Kb   114Kb
rails-app-4.localdomain             => ip-10-2-87-122.eu-west-1.compute.i  7.23Kb  32.7Kb  32.7Kb
                                    <=                                     43.0Kb   265Kb   265Kb
rails-app-4.localdomain             => ip-10-2-87-246.eu-west-1.compute.i  24.6Kb  33.7Kb  33.7Kb
                                    <=                                      191Kb   232Kb   232Kb
rails-app-4.localdomain             => ip-10-2-20-217.eu-west-1.compute.i  50.8Kb  56.5Kb  56.5Kb
                                    <=                                     81.7Kb  73.2Kb  73.2Kb
-------------------------------------------------------------------------------------------------
```

Resource: Network XI

```
TX:             cum:   1.77MB   peak:   4.23Mb                    rates:   4.23Mb  3.55Mb  3.55Mb
RX:                    5.41MB           16.6Mb                             16.6Mb  10.8Mb  10.8Mb
TOTAL:                 7.18MB           20.8Mb                             20.8Mb  14.4Mb  14.4Mb
```

- View socket connections

```
# ss
Netid  State      Recv-Q Send-Q Local Address:Port                 Peer Address:Port
u_str  ESTAB      0      0                * 655403                         * 655402
u_str  ESTAB      0      0      /tmp/passenger.RgJUBdx/apps.s/preloader.1uya910 81482655
u_str  ESTAB      0      0                * 81813975                       * 81813976
u_str  ESTAB      0      0                * 81803882                       * 81804840
........................
u_str  ESTAB      0      0      /tmp/passenger.RgJUBdx/apps.s/preloader.1uya910 81804840
u_str  ESTAB      0      0      /tmp/passenger.RgJUBdx/apps.s/ruby.JYtJfWTrsC07A8epKaneYMZ58GHSsR6NAhkVUkCKaeqa
u_str  ESTAB      0      0      /tmp/passenger.RgJUBdx/apps.s/preloader.1uya910 81813976
u_str  ESTAB      0      0      /tmp/passenger.RgJUBdx/apps.s/ruby.JX11S6SGXlpBiU6QLl2YgHjSHQwgmUHI8wdW0jXQBoNl
u_str  ESTAB      0      0      /tmp/passenger.RgJUBdx/apps.s/preloader.1uya910 81670404
u_str  ESTAB      0      0                * 81876765                       * 81874494
u_str  ESTAB      0      0      /tmp/passenger.RgJUBdx/apps.s/ruby.htx8W1Kaq8RwQ6KlCzhIyElbVvBUs4oNR6BPrJmsbCrJ
tcp    ESTAB      0      0      10.2.16.117:58804                50.31.164.148:http
tcp    ESTAB      0      0      10.2.16.117:60228                 10.2.86.42:mysql
tcp    LAST-ACK   0      32     10.2.16.117:54172                52.94.5.156:https
tcp    ESTAB      0      0      10.2.16.117:47040                10.2.20.217:6379
tcp    ESTAB      0      0      10.2.16.117:45890                10.2.20.217:6379
........................
tcp    ESTAB      0      0      10.2.16.117:59738                10.2.21.162:memcache
```

Resource: Network XII

```
tcp    ESTAB      0      0      10.2.16.117:45704                10.2.20.217:6379
tcp    ESTAB      0      0      10.2.16.117:46888                 10.2.20.64:memcache
........................
tcp    ESTAB      0      0      10.2.16.117:38134                50.31.164.149:http
tcp    ESTAB      0      0      10.2.16.117:45154                50.31.164.147:http
tcp    ESTAB      0      0      10.2.16.117:52644                10.2.204.15:http
tcp    ESTAB      0      0      10.2.16.117:41902                 10.2.20.76:6379
tcp    ESTAB      0      0      10.2.16.117:45784                10.2.20.217:6379
tcp    CLOSE-WAIT 1      0       ::ffff:10.2.16.117:46754                ::ffff:169.254.169.254:http
tcp    CLOSE-WAIT 32     0       ::ffff:10.2.16.117:47502                  ::ffff:176.34.227.36:https
tcp    ESTAB      0      0         ::ffff:127.0.0.1:31000                  ::ffff:127.0.0.1:32000
tcp    ESTAB      0      0       ::ffff:10.2.16.117:55954                  ::ffff:46.51.173.146:https

######################

You can get even more detailed information about a socket from kernel's internal socket struct.
For example, we will try to get nginx listening (on port 81) socket's backlog length

NOTE: You may need to install kernel debug info if not already installed
# yum-config-manager --enable "amzn-main-debuginfo" --enable "amzn-updates-debuginfo"
# yum -y install kernel-debuginfo kernel-devel

Or get socket info for listening socket on port 81
# ss -len | grep :81
tcp    LISTEN     0      511       *:81                    *:*                     ino:29842919 sk:55 <->

Get its sk buff address
# grep 29842919 /proc/net/tcp
   8: 00000000:0051 00000000:0000 0A 00000000:00000000 00:00000000 00000000     0        0 29842919 1 ffff8800e
```

Resource: Network XIII

```
Now get the details, for example, the backlog length
# gdb /usr/lib/debug/lib/modules/‘uname -r‘/vmlinux /proc/kcore
................
................
Reading symbols from /usr/lib/debug/lib/modules/4.4.51-40.69.amzn1.x86_64/vmlinux...done.
[New process 1]
Core was generated by ‘root=LABEL=/ console=tty1 console=ttyS0 selinux=0 LANG=en_US.UTF-8 KEYTABLE=us’.
................
(gdb) set print pretty on
(gdb) p *(struct sock *)0xffff8800e9d1da00
................
................
  sk_ack_backlog = 0,
  sk_max_ack_backlog = 511,
................
................
```

- Network latency / reachability

```
# Latency
# ping -c 3 google.com
PING google.com (172.217.7.142) 56(84) bytes of data.
64 bytes from iad30s08-in-f14.1e100.net (172.217.7.142): icmp_seq=1 ttl=48 time=1.57 ms
64 bytes from iad30s08-in-f142.1e100.net (172.217.7.142): icmp_seq=2 ttl=48 time=1.12 ms
64 bytes from iad30s08-in-f14.1e100.net (172.217.7.142): icmp_seq=3 ttl=48 time=1.13 ms

--- google.com ping statistics ---
3 packets transmitted, 3 received, 0% packet loss, time 2002ms
```

Resource: Network XIV

- Errors

```
# Look for errors, dropped, overruns etc

# ethtool -S eth0
NIC statistics:
     rx_packets: 604359552
     tx_packets: 649217118
     rx_bytes: 616073831008
     tx_bytes: 278008238445
     tx_busy: 0
     tx_restart_queue: 0
     tx_timeout_count: 0
     multicast: 0
     rx_csum_offload_errors: 0
     rx_bp_poll_yield: 0
     rx_bp_cleaned: 0
     rx_bp_misses: 0
     tx_bp_napi_yield: 0
     tx_bp_cleaned: 0
     tx_bp_misses: 0
```

- Looking at live network traffic

Resource: Network XV

```
Just display the TCP connection establishment alone
# tshark -f ’(tcp[tcpflags] & (tcp-syn) != 0)’
Running as user "root" and group "root". This could be dangerous.
Capturing on eth0
0.000000000 172.16.10.27 -> 172.16.17.173 TCP 74 40478 > 81 [SYN] Seq=0 Win=26883 Len=0 MSS=8961 SACK_PERM=1 TS
0.000015819 172.16.17.173 -> 172.16.10.27 TCP 74 81 > 40478 [SYN, ACK] Seq=0 Ack=1 Win=26847 Len=0 MSS=8961 SAC
0.063601284 172.16.10.153 -> 172.16.17.173 TCP 74 56494 > 81 [SYN] Seq=0 Win=26883 Len=0 MSS=8961 SACK_PERM=1 T
0.063618642 172.16.17.173 -> 172.16.10.153 TCP 74 81 > 56494 [SYN, ACK] Seq=0 Ack=1 Win=26847 Len=0 MSS=8961 SA
0.074333351 172.16.10.13 -> 172.16.17.173 TCP 74 37818 > 81 [SYN] Seq=0 Win=26883 Len=0 MSS=8961 SACK_PERM=1 TS
.................
.................
```

Resource: Various software resources I

- Global resource limits
- Process/thread specific resource limits
- cgroup
- Lock contention

```
# Certain resource limits may be set at system level as well as process level

# Maximum system level file descriptors
# sysctl fs.file-max fs.file-nr
fs.file-max = 762054
fs.file-nr = 1888       0       762054

# Maximum system level processes/threads
# sysctl kernel.threads-max kernel.pid_max
kernel.threads-max = 59690
kernel.pid_max = 32768

# Process level limits
# cat /proc/$(pidof ruby | awk '{print $1}')/limits
Limit                     Soft Limit           Hard Limit           Units
Max cpu time              unlimited            unlimited            seconds
Max file size             unlimited            unlimited            bytes
Max data size             unlimited            unlimited            bytes
Max stack size            8388608              unlimited            bytes
```

Resource: Various software resources II

```
Max core file size        0                    unlimited            bytes
Max resident set          unlimited            unlimited            bytes
Max processes             29845                29845                processes
Max open files            1024                 4096                 files
Max locked memory         65536                65536                bytes
Max address space         unlimited            unlimited            bytes
Max file locks            unlimited            unlimited            locks
Max pending signals       29845                29845                signals
Max msgqueue size         819200               819200               bytes
Max nice priority         0                    0
Max realtime priority     0                    0
Max realtime timeout      unlimited            unlimited            us

# ls -l /proc/$(pidof ruby | awk ’{print $1}’)/fd | wc -l
23

# Number of threads in a given process
# ls -l /proc/$(pidof ruby | awk ’{print $1}’)/task/ | wc -l
6

# Where/on what a given process is waiting on
# cat /proc/$(pidof ruby | awk ’{print $1}’)/stack
[<ffffffff811e8219>] poll_schedule_timeout+0x49/0x70
[<ffffffff811e8bac>] do_select+0x58c/0x750
[<ffffffff811e8f3c>] core_sys_select+0x1cc/0x2d0
[<ffffffff811e90eb>] SyS_select+0xab/0xf0
[<ffffffff814f002e>] entry_SYSCALL_64_fastpath+0x12/0x71
[<ffffffffffffffff>] 0xffffffffffffffff
```

Xen specific: Exploring xenstore from DomU I

```
# Install dependencies
$ yum groupinstall -y ’Development Tools’
$ yum install -y dev86 iasl ncurses-devel glib2-devel pixman-devel \
  libaio-devel glibc-devel.i686 cmake xz-devel libuuid-devel \
  zlib-devel
$ pushd ~
$ wget http://github.com/lloyd/yajl/tarball/2.1.0
$ tar xvf 2.1.0
$ cd lloyd-yajl-66cb08c/
$ ./configure
$ make
$ sudo make install
$ popd

# Install xen tools
$ wget https://downloads.xenproject.org/release/xen/4.9.0/xen-4.9.0.tar.gz
$ tar xvf xen-4.9.0.tar.gz
$ cd xen-4.9.0/tools/
$ ./configure
$ make -C include
$ make -C ./libs
$ make -C ./libxc
$ make -C ./xenstore
$ sudo make install -C ./xenstore bindir=/usr/local/bin libdir=/usr/local/lib
$ export PATH=$PATH:/usr/local/bin
$ export LD_LIBRARY_PATH=$LD_LIBRARY_PATH:/usr/local/lib
$ sudo ldconfig

$ sudo mount -t xenfs none /proc/xen
```

Xen specific: Exploring xenstore from DomU II

```
###########################################################

# List the store exposed to this DomU
$ sudo su
$ xenstore-ls /local/domain/‘xenstore-read domid‘
vm = "/vm/ec2a1431-14c4-fed6-dabd-eed158b16aa4"
device = ""
 vbd = ""
  51712 = ""
   backend-id = "0"
   virtual-device = "51712"
   device-type = "disk"
   state = "4"
   backend = "/local/domain/0/backend/vbd/204/51712"
   ring-ref = "8"
   event-channel = "35"
   protocol = "x86_64-abi"
   feature-persistent = "1"
  51824 = ""
   backend-id = "0"
   virtual-device = "51824"
   device-type = "disk"
   state = "4"
   backend = "/local/domain/0/backend/vbd/204/51824"
   ring-ref = "1090"
   event-channel = "40"
   protocol = "x86_64-abi"
   feature-persistent = "1"
 pci = ""
```

Xen specific: Exploring xenstore from DomU III

```
  0 = ""
   state = "1"
   backend-id = "0"
   backend = "/local/domain/0/backend/pci/204/0"
 console = ""
  0 = ""
   state = "1"
   backend-id = "0"
   backend = "/local/domain/0/backend/console/204/0"
control = ""
 platform-feature-multiprocessor-suspend = "1"
 platform-feature-xs_reset_watches = "1"
error = ""
memory = ""
 target = "7864320"
guest = ""
hvmpv = ""
data = ""
image = ""
 device-model-fifo = "/var/run/xend/dm-204-1502179628.fifo"
 device-model-pid = "11313"
 suspend-cancel = "1"
console = ""
 vnc-port = "5905"
 vnc-listen = "127.0.0.1"
 vnc-pass = "xyz"
 port = "7"
 limit = "1048576"
 type = "ioemu"
serial = ""
```

Xen specific: Exploring xenstore from DomU IV

```
 0 = ""
  tty = "/dev/pts/5"
description = ""
cpu = ""
 2 = ""
  availability = "online"
 0 = ""
  availability = "online"
 3 = ""
  availability = "online"
 1 = ""
  availability = "online"
domid = "204"
store = ""
 ring-ref = "1044476"
 port = "6"
name = "dom_27677293465"
device-misc = ""
 console = ""
  nextDeviceID = "1"

# List one specific VBD device from above
$ xenstore-ls  /local/domain/0/backend/vbd/826/51728
domain = "dom_24698651860"
frontend = "/local/domain/826/device/vbd/51728"
uuid = "d62daa8e-d864-e843-f455-3640ffc3cfbf"
bootable = "0"
dev = "xvdb"
state = "4"
params = "/dev/nvme2n1"
```

Xen specific: Exploring xenstore from DomU V

```
mode = "w"
removable = "1"
online = "1"
frontend-id = "826"
type = "phy"
physical-device = "fb:40"
hotplug-status = "connected"
feature-flush-cache = "0"
feature-discard = "0"
feature-barrier = "0"
feature-persistent = "1"
feature-max-indirect-segments = "256"
sectors = "209715200"
info = "0"
sector-size = "512"
physical-sector-size = "512"
```

Topic

1. System
2. Application
3. Profilers
4. Debuggers
5. Books

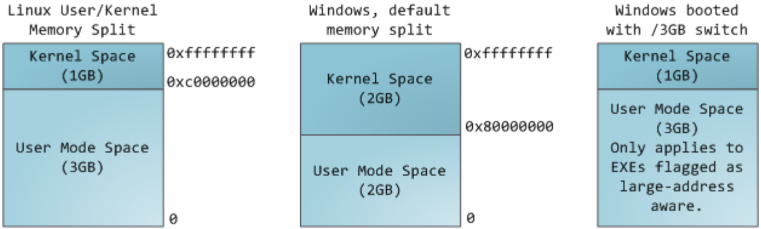
Application performance analysis

- Resource based USE method is what we saw so far for identifying system level performance issues.
- For identifying application performance issues, we could identify what the application is doing over a period then analzye: TSA (Thread State Analysis) method [7] is something we could use.
- For the application/process we want to analyze, identify the threads it has
- For each thread
  - Measure time spent in each state
    - State can be R (running), S (sleeping), D (uninterruptible sleep), T (stopped), t (stopped by debugger), Z (zombie)
  - Investigate states from most frequent to least
- But before we introduce/use various tools to do that, we have to know various things about process, so we will take a detour and do a deep dive into process

[7] `http://www.brendangregg.com/tsamethod.html`

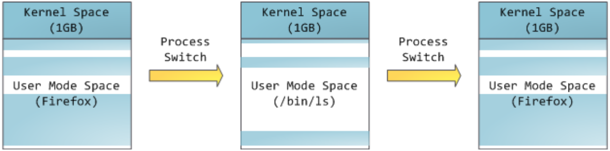
Process I

- Introduction: http://duartes.org/gustavo/blog/post/anatomy-of-a-program-in-memory/
- User Space vs Kernel Space split (32bit OS)


Linux User/Kernel Memory Split
Kernel Space (1GB)
User Mode Space (3GB)
0xffffffff
0xc0000000
0
Windows, default memory split
Kernel Space (2GB)
User Mode Space (2GB)
0xffffffff
0x80000000
0
Windows booted with /3GB switch
Kernel Space (1GB)
User Mode Space (3GB) Only applies to EXEs flagged as large-address aware.

Process II

- Kernel Space is same across processes, only User Space content changes across processes



- Process is always started from an executable binary file in certain format (ELF is the most common in Linux). That format defines/standardizes various things that would be read/used by OS during process creation
  - A typical example, for Golang: Your source code is compiled into native machine code and an ELF file is created with all the info, including the generated machine code, that OS can use during process creation.
  - Note that scripted applications, like Ruby, first a process (VM, generally written in C/C++) is started, which has logic to parse/run the ruby scripts further.

Process III

- The executable binary may be
  - Statically linked: All dependent code/data is included in the binary and is self contained
  - Or dynamically linked: Certain code/data it depends on comes from another binary (typically a shared library .so) and will only be resolved during process startup. The tool that does "runtime resolving" is called a "loader".
- Let's take a look at how the executable looks on disk.
  We will use ruby executable as an example
  - What kind of file it is?

```
$ file /usr/local/bin/ruby
/usr/local/bin/ruby: ELF 64-bit LSB executable, x86-64, version 1 (SYSV), dynamically linked,
   interpreter /lib64/ld-linux-x86-64.so.2, for GNU/Linux 2.6.35,
   BuildID[sha1]=27875858789fb14bfbf4ac2c603ec700acff91da, not stripped
```

    It is 64 bit ELF binary, dynamically linked, using loader at /lib64/ld-linux-x86-64.so.2 and contains debug symbols
  - What it depends on (shared libraries)?

Process IV

```
$ ldd /usr/local/bin/ruby
        linux-vdso.so.1 =>  (0x00007ffebc99c000)
        libruby.so.2.2 => /usr/local/lib/libruby.so.2.2 (0x00007fbb49095000)
        libpthread.so.0 => /lib64/libpthread.so.0 (0x00007fbb48e79000)
        libdl.so.2 => /lib64/libdl.so.2 (0x00007fbb48c74000)
        libcrypt.so.1 => /lib64/libcrypt.so.1 (0x00007fbb48a3d000)
        libm.so.6 => /lib64/libm.so.6 (0x00007fbb4873b000)
        libc.so.6 => /lib64/libc.so.6 (0x00007fbb48378000)
        /lib64/ld-linux-x86-64.so.2 (0x00005634741af000)
        libfreebl3.so => /lib64/libfreebl3.so (0x00007fbb48176000)
```

- What it contains?

```
$ size /usr/local/bin/ruby
   text    data     bss     dec     hex filename
   2298     668       4    2970     b9a /usr/local/bin/ruby

# Or all sections
$ size -Ax /usr/local/bin/ruby
/usr/local/bin/ruby  :
section                 size       addr
.interp                 0x1c   0x400200
.note.ABI-tag           0x20   0x40021c
.note.gnu.build-id      0x24   0x40023c
.gnu.hash               0x50   0x400260
.dynsym                0x228   0x4002b0
.dynstr                0x164   0x4004d8
.gnu.version            0x2e   0x40063c
.gnu.version_r          0x20   0x400670
.rela.dyn               0x18   0x400690
```

Process V

```
.rela.plt               0xc0   0x4006a8
.init                   0x1a   0x400768
.plt                    0x90   0x400790
.text                  0x1b4   0x400820
.fini                    0x9   0x4009d4
.rodata                 0x11   0x4009e0
.eh_frame_hdr           0x34   0x4009f4
.eh_frame               0xec   0x400a28
.init_array              0x8   0x600b18
.fini_array              0x8   0x600b20
.jcr                     0x8   0x600b28
.dynamic               0x220   0x600b30
.got                     0x8   0x600d50
.got.plt                0x58   0x600d58
.data                    0x4   0x600db0
.bss                     0x4   0x600db4
.comment                0x2c        0x0
.debug_aranges          0x30        0x0
.debug_info            0x492        0x0
.debug_abbrev          0x18d        0x0
.debug_line            0x538        0x0
.debug_str           0x194f1        0x0
.debug_loc              0xaa        0x0
.debug_ranges           0x20        0x0
.debug_macro          0x56ac        0x0
Total                0x203b4
```

- How does ELF format look like?

Process VI

```
# Show the header
$ readelf -h /usr/local/bin/ruby
ELF Header:
  Magic:   7f 45 4c 46 02 01 01 00 00 00 00 00 00 00 00 00
  Class:                             ELF64
  Data:                              2's complement, little endian
  Version:                           1 (current)
  OS/ABI:                            UNIX - System V
  ABI Version:                       0
  Type:                              EXEC (Executable file)
  Machine:                           Advanced Micro Devices X86-64
  Version:                           0x1
  Entry point address:               0x400870
  Start of program headers:          64 (bytes into file)
  Start of section headers:          132928 (bytes into file)
  Flags:                             0x0
  Size of this header:               64 (bytes)
  Size of program headers:           56 (bytes)
  Number of program headers:         8
  Size of section headers:           64 (bytes)
  Number of section headers:         38
  Section header string table index: 35

# Show various sections
$ readelf -S /usr/local/bin/ruby
There are 38 section headers, starting at offset 0x20740:

Section Headers:
  [Nr] Name              Type             Address           Offset
       Size              EntSize          Flags  Link  Info  Align
```

Process VII

```
  [ 0]                   NULL             0000000000000000  00000000
       0000000000000000  0000000000000000           0     0     0
  [ 1] .interp           PROGBITS         0000000000400200  00000200
       000000000000001c  0000000000000000   A       0     0     1
  .....................
  .....................
  [13] .text             PROGBITS         0000000000400820  00000820
       00000000000001b4  0000000000000000  AX       0     0     16
  [14] .fini             PROGBITS         00000000004009d4  000009d4
       0000000000000009  0000000000000000  AX       0     0     4
  [15] .rodata           PROGBITS         00000000004009e0  000009e0
       0000000000000011  0000000000000000   A       0     0     8
  .....................
  .....................
  [28] .debug_info       PROGBITS         0000000000000000  00000e10
       0000000000000492  0000000000000000           0     0     1
  [29] .debug_abbrev     PROGBITS         0000000000000000  000012a2
       000000000000018d  0000000000000000           0     0     1
  .....................
  .....................
Key to Flags:
  W (write), A (alloc), X (execute), M (merge), S (strings), l (large)
  I (info), L (link order), G (group), T (TLS), E (exclude), x (unknown)
  O (extra OS processing required) o (OS specific), p (processor specific)


# Show the segment header (i.e., on memory layout)
$ readelf -l ‘which ruby‘

Elf file type is EXEC (Executable file)
```

Process VIII

```
Entry point 0x400870
There are 8 program headers, starting at offset 64

Program Headers:
  Type           Offset             VirtAddr           PhysAddr
                 FileSiz            MemSiz              Flags  Align
  PHDR           0x0000000000000040 0x0000000000400040 0x0000000000400040
                 0x00000000000001c0 0x00000000000001c0  R E    8
  INTERP         0x0000000000000200 0x0000000000400200 0x0000000000400200
                 0x000000000000001c 0x000000000000001c  R      1
      [Requesting program interpreter: /lib64/ld-linux-x86-64.so.2]
  LOAD           0x0000000000000000 0x0000000000400000 0x0000000000400000
                 0x0000000000000b14 0x0000000000000b14  R E    200000
  LOAD           0x0000000000000b18 0x0000000000600b18 0x0000000000600b18
                 0x000000000000029c 0x00000000000002a0  RW     200000
  DYNAMIC        0x0000000000000b30 0x0000000000600b30 0x0000000000600b30
                 0x0000000000000220 0x0000000000000220  RW     8
  NOTE           0x000000000000021c 0x000000000040021c 0x000000000040021c
                 0x0000000000000044 0x0000000000000044  R      4
  GNU_EH_FRAME   0x00000000000009f4 0x00000000004009f4 0x00000000004009f4
                 0x0000000000000034 0x0000000000000034  R      4
  GNU_STACK      0x0000000000000000 0x0000000000000000 0x0000000000000000
                 0x0000000000000000 0x0000000000000000  RW     10

 Section to Segment mapping:
  Segment Sections...
   00
   01     .interp
   02     .interp .note.ABI-tag .note.gnu.build-id .gnu.hash .dynsym .dynstr .gnu.version .gnu.ve
   03     .init_array .fini_array .jcr .dynamic .got .got.plt .data .bss
```

Process IX

```
   04     .dynamic
   05     .note.ABI-tag .note.gnu.build-id
   06     .eh_frame_hdr
   07

# Or use objdump to see them together
$ objdump -h /usr/local/bin/ruby

/usr/local/bin/ruby:     file format elf64-x86-64

Sections:
Idx Name          Size      VMA               LMA               File off  Algn
  0 .interp       0000001c  0000000000400200  0000000000400200  00000200  2**0
                  CONTENTS, ALLOC, LOAD, READONLY, DATA
...............
...............
 11 .plt          00000090  0000000000400790  0000000000400790  00000790  2**4
                  CONTENTS, ALLOC, LOAD, READONLY, CODE
 12 .text         000001b4  0000000000400820  0000000000400820  00000820  2**4
                  CONTENTS, ALLOC, LOAD, READONLY, CODE
 13 .fini         00000009  00000000004009d4  00000000004009d4  000009d4  2**2
                  CONTENTS, ALLOC, LOAD, READONLY, CODE
 14 .rodata       00000011  00000000004009e0  00000000004009e0  000009e0  2**3
                  CONTENTS, ALLOC, LOAD, READONLY, DATA
...............
...............
 23 .data         00000004  0000000000600db0  0000000000600db0  00000db0  2**2
                  CONTENTS, ALLOC, LOAD, DATA
 24 .bss          00000004  0000000000600db4  0000000000600db4  00000db4  2**2
                  ALLOC
```

Process X

```
...............
...............
 25 .comment      0000002c  0000000000000000  0000000000000000  00000db4  2**0
                  CONTENTS, READONLY
 26 .debug_aranges 00000030  0000000000000000  0000000000000000  00000de0  2**0
                  CONTENTS, READONLY, DEBUGGING
...............
...............
 33 .debug_macro  000056ac  0000000000000000  0000000000000000  0001af22  2**0
                  CONTENTS, READONLY, DEBUGGING
```

- Where is my function in it?

```
# Use nm to look for symbols, for example, the entry point 0x400870 we found above,
# we can check what function is contained there.
$ nm /usr/local/bin/ruby
0000000000600db4 B __bss_start
...............
...............
0000000000400820 t main
00000000004008d0 t register_tm_clones
                 U ruby_init
                 U ruby_init_stack
                 U ruby_options
                 U ruby_run_node
                 U ruby_sysinit
                 U setlocale@@GLIBC_2.2.5
0000000000400870 T _start
...............
...............
```

Process XI

- How a process is started
  You can find a detailed explanation here [8]
  - Some process wants to start a new program (say, shell, wants to start /usr/local/bin/ruby program), so it calls into kernel (using fork/exec)
  - Kernel checks what type of file it is (binfmt kernel feature)
    - It could be ELF, java binary, .net binary, shell script with shebang "#!/bin/sh" etc
    - Linux has extensible support via binfmt [9]
  - If it is a supported file, in this case ELF executable, it will load it and pass control to [10]
    - "Load" == finding various ELF sections we saw above, memory mapping them
    - "Pass" == passing execution control to the entry point specified in the ELF binary
  - Additionally what executable gets loaded/run may vary based whether the executable is statically linked or dynamically linked
  - If it is dynamically linked

Process XII

- Instead of loading the executable, it will load the loader (/lib64/ld-linux-x86-64.so.2) and pass control to it, along with info about the file to be exectued (via AUX info). Note that loader is just another normal executable as far as kernel is concerned.

```
$ LD_SHOW_AUXV=1 /usr/local/bin/ruby
AT_SYSINFO_EHDR: 0x7ffdf92cf000
AT_HWCAP:        178bfbff
AT_PAGESZ:       4096
AT_CLKTCK:       100
AT_PHDR:         0x400040
AT_PHENT:        56
AT_PHNUM:        8
AT_BASE:         0x7f32eb690000
AT_FLAGS:        0x0
AT_ENTRY:        0x400870
AT_UID:          2147
AT_EUID:         2147
AT_GID:          501
AT_EGID:         501
AT_SECURE:       0
AT_RANDOM:       0x7ffdf928b049
AT_EXECFN:       /usr/local/bin/ruby
AT_PLATFORM:     x86_64
```

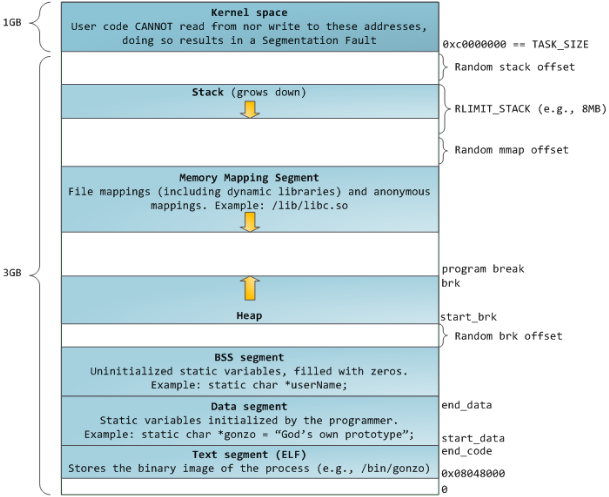
- Loader in turn loads the /usr/local/bin/ruby executable and passes control to the entry point specified

Process XIII

  - If it is statically linked, kernel loads the executable and passes control to the entry point specified
  - Note that in either case (statically or dynamically linked), as far as kernel is concerned, it is just going to load an executable and pass control to it. In the case of dynamically linked executable, it just happens to be the "loader" executable.
  - Additionally, other executables, like Java etc can be started via similar method (i.e., kernel loads the JVM and passes control to it etc)
  - I have a detailed info here that covers execve syscall + loader initialization + symbol resolving + binary execution
- Let's take a look at how it looks at runtime

Process XIV

- A typical memory layout of a process



1GB

Kernel space
User code CANNOT read from nor write to these addresses, doing so results in a Segmentation Fault

0xc0000000 == TASK_SIZE

Random stack offset

Stack (grows down)

RLIMIT_STACK (e.g., 8MB)

Random mmap offset

Memory Mapping Segment
File mappings (including dynamic libraries) and anonymous mappings. Example: /lib/libc.so

3GB

program break
brk

Heap

start_brk

Random brk offset

BSS segment
Uninitialized static variables, filled with zeros.
Example: static char *userName;

end_data

Data segment
Static variables initialized by the programmer.
Example: static char *gonzo = “God’s own prototype”;

start_data
end_code

Text segment (ELF)
Stores the binary image of the process (e.g., /bin/gonzo)

0x08048000

0

- How do I see a process’s memory layout?

Process XV

```
cat /proc/19206/maps
00400000-00401000 r-xp 00000000 ca:01 29854                              /usr/local/bin/ruby
00600000-00601000 rw-p 00000000 ca:01 29854                              /usr/local/bin/ruby
01cca000-1f6ca000 rw-p 00000000 00:00 0                                  [heap]
1f6ca000-4123c000 rw-p 00000000 00:00 0                                  [heap]
7f21ef922000-7f21ef923000 ---p 00000000 00:00 0
7f21ef923000-7f21efa23000 rw-p 00000000 00:00 0
7f21efa23000-7f21efa24000 ---p 00000000 00:00 0
7f21efa24000-7f21efb24000 rw-p 00000000 00:00 0
7f21efb24000-7f21efb44000 r-xp 00000000 ca:01 3880                       /usr/lib64/libnssdbm3.so
...............
...............
7f21f10f7000-7f21f10fb000 r-xp 00000000 ca:70 524440                     /data/helpkit/shared/bun
7f21fc0da000-7f21fc2d9000 ---p 00258000 ca:01 285180                     /usr/lib64/mysql/libmysq
...............
...............
7f2209462000-7f2209726000 r-xp 00000000 ca:01 29858                      /usr/local/lib/libruby.s
7f2209726000-7f2209925000 ---p 002c4000 ca:01 29858                      /usr/local/lib/libruby.s
...............
...............
7f2209baf000-7f2209bcf000 r-xp 00000000 ca:01 268462                     /lib64/ld-2.17.so
...............
...............
7fffb3d23000-7fffb4522000 rw-p 00000000 00:00 0                          [stack]
7fffb4537000-7fffb4539000 r--p 00000000 00:00 0                          [vvar]
7fffb4539000-7fffb453b000 r-xp 00000000 00:00 0                          [vdso]
ffffffffff600000-ffffffffff601000 r-xp 00000000 00:00 0                  [vsyscall]
```

- Stack

Process XVI

- Heap
  - Where all the dynamic variables/objects from the process are stored
  - How it is managed: Manually or automatically
- Virtual vs Resident
  - Resident: RSS vs PSS vs USS

```
# Check memory usage, system-wide
$ sudo smem -k
  PID User     Command                         Swap      USS      PSS      RSS
 2972 root     /sbin/mingetty /dev/tty4           0    88.0K   110.0K     1.4M
.......
.......
22190 deploy   Passenger AppPreloader: /da        0   141.4M   196.6M   509.4M
22800 deploy   Passenger RubyApp: /data/he        0   160.8M   213.4M   520.7M
22788 deploy   Passenger RubyApp: /data/he        0   161.1M   213.5M   520.6M
22772 deploy   Passenger RubyApp: /data/he        0   161.4M   213.8M   520.9M
22744 deploy   Passenger RubyApp: /data/he        0   177.1M   226.9M   526.0M
22754 deploy   Passenger RubyApp: /data/he        0   177.1M   226.9M   526.3M
18599 root     /opt/SumoCollector/jre/bin/        0   233.1M   233.2M   235.3M
22731 deploy   Passenger RubyApp: /data/he        0   188.4M   237.3M   526.6M
```

- 'Thread': Heap is shared, stack is unique to each 'thread'
- Process state (R, S, D etc)
- Tools: top, htop, pmap, smem etc

[8] https://github.com/0xAX/linux-insides/blob/master/Misc/program_startup.md

[9] https://en.wikipedia.org/wiki/Binfmt_misc

[10] https://github.com/torvalds/linux/blob/v4.13/fs/binfmt_elf.c#L679

Process runtime I

- Some processes may have a runtime and some may not have, based on what kind language they were built with
  - Minimal or no runtime: C/C++, Rust etc
  - With runtime: Java/JVM, C#/CLR, Go, Ruby, Python etc
- How they execute code
  - Compiled to native: C/C++, Go, Rust, C# + ngen (AOT), Java + AOT etc
  - Interpreted: MRI Ruby, Python, Perl, Erlang/BeamVM, Node.js, Java etc
    - Some of them may have intermediate form, but they can still be interepreted: Example, Java => bytecode => Interepreted
  - JITed: C#, Java, Node.js
    - Some may alternate between interpreted mode and JIT (example: Java, node.js)
    - Some always start in JITed mode: CLR/C#
- By how they manage memory [11]
  - Manual memory management: C/C++ etc

Process runtime II

  - Automatic memory management
    - Garbage collection: Java, Go, Ruby, Python etc
    - Reference counting: Objective-C, Python, Rust, C++
    - Resource Acquisition Is Initialization (RAII): Rust, C++
- By how they manage concurrency/parallelism (== threading)
  - Single threading
  - Multi threading
    - 1:1 threading: C, C++, Java, Ruby, Python etc
    - N:1 thread: Ruby fibers [12], [13]
    - M:N threading: Golang, BeamVM (Erlang, Elixir)
  - Multi-process model
  - Evented vs Threaded
    - How blocking operations are handled
    - Nodejs, Go example
- Examples:
  - Single threaded: MRI Ruby
  - Evented: nginx, haproxy, nodejs, Ruby + EventMachine

Process runtime III

- Evented + multi-threaded: golang, nodejs: blocking operations are sent to thread pool
- Multi-process + evented: nginx, haproxy

[11] https://www.cs.virginia.edu/~cs415/reading/bacon-garbage.pdf

[12] http://schmurfy.github.io/2011/09/25/on_fibers_and_threads.html

[13] http://oldmoe.blogspot.in/2008/08/ruby-fibers-vs-ruby-threads.html

System Calls I

- Introduction to syscalls [14] , [15]
- What is syscall: User Mode code requesting a service from Kernel Mode. Example: Writing to a file, sending a data out over network etc

Table: List of typical syscalls

syscall	what it does
open	Open a file, returns file descriptor
socket	Open a socket, returns file descriptor
read	Read data from a file descriptor (file, socket)
write	Write data to a file descriptor (file, socket)
close	Close a file descriptor
fork	Create a new process (out of current process)
exec/execve	Replace current process with new program
connect	Connect to remote host
accept	Accept a new connection on a socket
stat	Get file status
ioctl	Perform control functions on file descriptor
mmap	Map a file to the process address space
brk	Extend the heap pointer

System Calls II

- Blocking vs Non-Blocking
  - Not all syscalls have non-blocking option
- Since syscall is the primary way processes interact with system, finding out what kind of syscalls a process is making could give us very good insight into what it is doing: This can be used for debugging, profiling or just for general understanding of a given process.
- How to find what kind of syscalls a process is making?: 'strace' or 'perf trace'
- How strace is implemented
  - Using ptrace interface [16]
  - This will cause two context switches for each syscall traced, can slow down the program significantly if it uses too many syscalls.
  - 'perf trace' is better in terms of performance, use that over strace if possible
- Things to remember

System Calls III

  - Performance impact of tracing in production: don't use it unless really required
  - Underlying language 'runtime' semantics (evented, threaded: 1:1 threading or M:N threading model etc)

- Examples:

```
# Summarize system calls made
# strace -c ruby -e 'puts "hello world"'
hello world
% time     seconds  usecs/call     calls    errors syscall
------ ----------- ----------- --------- --------- ----------------
 19.30    0.000183           1       191        97 open
 15.61    0.000148           0       408           lstat
 14.35    0.000136           1       115           read
  6.96    0.000066           2        42           brk
  5.91    0.000056           2        32           mmap
  5.80    0.000055           1        97           fstat
  5.38    0.000051           1        42           fcntl
  4.11    0.000039           0        96           close
  2.85    0.000027           1        44        41 ioctl
  2.53    0.000024           1        35           geteuid
  2.11    0.000020           1        23           mprotect
  2.00    0.000019           1        34           getuid
  2.00    0.000019           1        35           getegid
  1.90    0.000018           1        18           rt_sigaction
  1.69    0.000016          16         1           clone
```

System Calls IV

```
  1.69    0.000016           0        34           getgid
  1.37    0.000013           7         2           pipe2
  0.84    0.000008           3         3           getpid
  0.63    0.000006           2         3           prlimit64
  0.63    0.000006           3         2           getrandom
  0.42    0.000004           0        31         4 stat
  0.42    0.000004           2         2           futex
  0.42    0.000004           4         1           sched_getaffinity
  0.42    0.000004           4         1           clock_gettime
  0.32    0.000003           1         3           rt_sigprocmask
  0.21    0.000002           2         1           sigaltstack
  0.11    0.000001           0         8           lseek
  0.00    0.000000           0         2           write
  0.00    0.000000           0         3           munmap
  0.00    0.000000           0         1         1 access
  0.00    0.000000           0         1           execve
  0.00    0.000000           0         6           getdents
  0.00    0.000000           0         1           arch_prctl
  0.00    0.000000           0         1           set_tid_address
  0.00    0.000000           0         1           set_robust_list
------ ----------- ----------- --------- --------- ----------------
100.00    0.000948                  1320       143 total

########################################################

# Or use 'perf trace'
# perf trace -s ruby -e 'puts "hello world"'
hello world

 Summary of events:
```

System Calls V

```
ruby (18436), 10 events, 0.4%

  syscall            calls    total       min       avg       max      stddev
                             (msec)    (msec)    (msec)    (msec)       (%)
  --------------- -------- --------- --------- --------- ---------     ------
  read                   2     0.004     0.002     0.002     0.003     21.82%
  close                  2     0.004     0.002     0.002     0.002     11.40%
  poll                   1     0.000     0.000     0.000     0.000      0.00%


ruby (18435), 2638 events, 98.7%

  syscall            calls    total       min       avg       max      stddev
                             (msec)    (msec)    (msec)    (msec)       (%)
  --------------- -------- --------- --------- --------- ---------     ------
  lstat                408     1.215     0.002     0.003     0.009      1.69%
  open                 191     1.001     0.002     0.005     0.017      4.52%
  read                 115     0.620     0.002     0.005     0.040      7.77%
  clone                  1     0.315     0.315     0.315     0.315      0.00%
  brk                   42     0.270     0.001     0.006     0.014      7.19%
  stat                  31     0.177     0.002     0.006     0.010      8.57%
  close                 96     0.173     0.001     0.002     0.004      2.60%
  fstat                 97     0.170     0.001     0.002     0.003      2.50%
  mmap                  32     0.143     0.002     0.004     0.009      5.93%
  mprotect              23     0.111     0.003     0.005     0.008      6.21%
  ioctl                 44     0.081     0.001     0.002     0.007      6.61%
  fcntl                 42     0.068     0.001     0.002     0.003      3.13%
  futex                  2     0.056     0.003     0.028     0.053     89.36%
  getuid                34     0.053     0.001     0.002     0.004      5.53%
```

System Calls VI

```
   geteuid               35     0.048     0.001     0.001     0.004      5.22%
   getegid               35     0.046     0.001     0.001     0.002      2.88%
   getgid                34     0.045     0.001     0.001     0.002      3.10%
   getdents               6     0.041     0.001     0.007     0.013     31.37%
   munmap                 3     0.032     0.007     0.011     0.013     16.94%
   rt_sigaction          18     0.025     0.001     0.001     0.002      4.22%
   lseek                  8     0.014     0.001     0.002     0.002      6.17%
   write                  2     0.011     0.002     0.006     0.009     57.99%
   pipe2                  2     0.011     0.003     0.005     0.008     51.70%
   getrandom              2     0.007     0.003     0.003     0.004     22.68%
   rt_sigprocmask         3     0.005     0.001     0.002     0.003     26.26%
   access                 1     0.005     0.005     0.005     0.005      0.00%
   prlimit64              3     0.004     0.001     0.001     0.001      3.48%
   getpid                 3     0.004     0.001     0.001     0.001      0.35%
   clock_gettime          1     0.003     0.003     0.003     0.003      0.00%
   sched_getaffinity        1     0.002     0.002     0.002     0.002      0.00%
   arch_prctl             1     0.002     0.002     0.002     0.002      0.00%
   sigaltstack            1     0.001     0.001     0.001     0.001      0.00%
   set_robust_list        1     0.001     0.001     0.001     0.001      0.00%
   set_tid_address        1     0.001     0.001     0.001     0.001      0.00%

########################################################

# Trace all syscalls
strace -ttT -ff ruby -e 'puts "hello world"'

11:15:42.151594 execve("/home/suresh/.rvm/rubies/ruby-2.4.1/bin/ruby", ["ruby", "-e", "puts \"hello
   world\""], [/* 119 vars */]) = 0 <0.000155>
11:15:42.151909 brk(NULL)               = 0x2214000 <0.000007>
11:15:42.151955 access("/etc/ld.so.preload", R_OK) = -1 ENOENT (No such file or directory) <0.000009>
```

System Calls VII

```
11:15:42.152005 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/tls/x86_64/libruby.so.2.4", O_RDONLY|O_CLOEXEC)
   = -1 ENOENT (No such file or directory) <0.000015>
11:15:42.152047 stat("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/tls/x86_64", 0x7fff6893f510) = -1 ENOENT
   (No such file or directory) <0.000009>
11:15:42.152080 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/tls/libruby.so.2.4", O_RDONLY|O_CLOEXEC)
   = -1 ENOENT (No such file or directory) <0.000009>
11:15:42.152113 stat("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/tls", 0x7fff6893f510) = -1 ENOENT (No such
   file or directory) <0.000010>
11:15:42.152148 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/x86_64/libruby.so.2.4", O_RDONLY|O_CLOEXEC)
   = -1 ENOENT (No such file or directory) <0.000010>
11:15:42.152181 stat("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/x86_64", 0x7fff6893f510) = -1 ENOENT
   (No such file or directory) <0.000015>
11:15:42.152234 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/libruby.so.2.4", O_RDONLY|O_CLOEXEC) =
   3 <0.000016>
11:15:42.152277 read(3, "\177ELF\2\1\1\0\0\0\0\0\0\0\0\0\3\0>\0\1\0\0\0\360\335\2\0\0\0\0\0"..., 832)
   = 832 <0.000009>
11:15:42.152318 fstat(3, {st_mode=S_IFREG|0755, st_size=4794408, ...}) = 0 <0.000008>
11:15:42.152349 mmap(NULL, 8192, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) = 0x7f5124a2c000
   <0.000009>
11:15:42.152380 mmap(NULL, 5339072, PROT_READ|PROT_EXEC, MAP_PRIVATE|MAP_DENYWRITE, 3, 0) = 0x7f51242f3000
   <0.000007>
11:15:42.152404 mprotect(0x7f51245f0000, 2097152, PROT_NONE) = 0 <0.000010>
11:15:42.152451 mmap(0x7f51247f0000, 32768, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_FIXED|MAP_DENYWRITE,
   3, 0x2fd000) = 0x7f51247f0000 <0.000016>
11:15:42.152498 mmap(0x7f51247f8000, 75712, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_FIXED|MAP_ANONYMOUS,
   -1, 0) = 0x7f51247f8000 <0.000011>
11:15:42.152544 close(3)                = 0 <0.000006>
11:15:42.152587 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/libpthread.so.0", O_RDONLY|O_CLOEXEC) =
   -1 ENOENT (No such file or directory) <0.000012>
.................
```

System Calls VIII

```
11:15:42.166524 clone(strace: Process 5106 attached
child_stack=0x7f5124a2aff0,
   flags=CLONE_VM|CLONE_FS|CLONE_FILES|CLONE_SIGHAND|CLONE_THREAD|CLONE_SYSVSEM|CLONE_SETTLS|CLONE_PARENT_SETTI
   parent_tidptr=0x7f5124a2b9d0, tls=0x7f5124a2b700, child_tidptr=0x7f5124a2b9d0) = 5106 <0.000042>
[pid  5106] 11:15:42.166585 set_robust_list(0x7f5124a2b9e0, 24 <unfinished ...>
[pid  5087] 11:15:42.166603 getpid( <unfinished ...>
[pid  5106] 11:15:42.166615 <... set_robust_list resumed> ) = 0 <0.000016>
[pid  5087] 11:15:42.166625 <... getpid resumed> ) = 5087 <0.000016>
[pid  5106] 11:15:42.166635 prctl(PR_SET_NAME, "ruby-timer-thr") = 0 <0.000009>
[pid  5106] 11:15:42.166669 poll([{fd=3, events=POLLIN}, {fd=5, events=POLLIN}], 2, -1 <unfinished ...>
[pid  5087] 11:15:42.167364 geteuid()   = 1000 <0.000008>
[pid  5087] 11:15:42.167410 getegid()   = 100 <0.000010>
[pid  5087] 11:15:42.167897 brk(0x2343000) = 0x2343000 <0.000014>
[pid  5087] 11:15:42.168282 getuid()    = 1000 <0.000012>
[pid  5087] 11:15:42.168349 geteuid()   = 1000 <0.000013>
[pid  5087] 11:15:42.168397 getgid()    = 100 <0.000011>
[pid  5087] 11:15:42.168432 getegid()   = 100 <0.000009>
[pid  5087] 11:15:42.168613 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/site_ruby/2.4.0/enc/encdb.so",
   O_RDONLY|O_NONBLOCK|O_CLOEXEC) = -1 ENOENT (No such file or directory) <0.000022>
[pid  5087] 11:15:42.168691
   open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/site_ruby/2.4.0/x86_64-linux/enc/encdb.so",
   O_RDONLY|O_NONBLOCK|O_CLOEXEC) = -1 ENOENT (No such file or directory) <0.000012>
[pid  5087] 11:15:42.168750 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/site_ruby/enc/encdb.so",
   O_RDONLY|O_NONBLOCK|O_CLOEXEC) = -1 ENOENT (No such file or directory) <0.000015>
[pid  5087] 11:15:42.168796
   open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/vendor_ruby/2.4.0/enc/encdb.so",
   O_RDONLY|O_NONBLOCK|O_CLOEXEC) = -1 ENOENT (No such file or directory) <0.000015>
[pid  5087] 11:15:42.168848
   open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/vendor_ruby/2.4.0/x86_64-linux/enc/encdb.so",
   O_RDONLY|O_NONBLOCK|O_CLOEXEC) = -1 ENOENT (No such file or directory) <0.000014>
```

System Calls IX

```
[pid  5087] 11:15:42.168901 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/vendor_ruby/enc/encdb.so",
   O_RDONLY|O_NONBLOCK|O_CLOEXEC) = -1 ENOENT (No such file or directory) <0.000015>
[pid  5087] 11:15:42.168950 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/2.4.0/enc/encdb.so",
   O_RDONLY|O_NONBLOCK|O_CLOEXEC) = -1 ENOENT (No such file or directory) <0.000011>
[pid  5087] 11:15:42.168999
   open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/2.4.0/x86_64-linux/enc/encdb.so",
   O_RDONLY|O_NONBLOCK|O_CLOEXEC) = 8 <0.000016>
[pid  5087] 11:15:42.169049 fcntl(8, F_GETFD) = 0x1 (flags FD_CLOEXEC) <0.000008>
[pid  5087] 11:15:42.169081 fstat(8, {st_mode=S_IFREG|0755, st_size=90384, ...}) = 0 <0.000007>
[pid  5087] 11:15:42.169120 close(8)    = 0 <0.000010>
[pid  5087] 11:15:42.169210 futex(0x7f5123e41048, FUTEX_WAKE_PRIVATE, 2147483647) = 0 <0.000008>
[pid  5087] 11:15:42.169247
   open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/2.4.0/x86_64-linux/enc/encdb.so", O_RDONLY|O_CLOEXEC)
   = 8 <0.000014>
[pid  5087] 11:15:42.169283 read(8, "\177ELF\2\1\1\0\0\0\0\0\0\0\0\0\3\0>\0\1\0\0\0\360\7\0\0\0\0\0\0"...,
   832) = 832 <0.000008>
[pid  5087] 11:15:42.169318 fstat(8, {st_mode=S_IFREG|0755, st_size=90384, ...}) = 0 <0.000007>
[pid  5087] 11:15:42.169353 mmap(NULL, 2109536, PROT_READ|PROT_EXEC, MAP_PRIVATE|MAP_DENYWRITE, 8, 0)
   = 0x7f5122f69000 <0.000015>
[pid  5087] 11:15:42.169392 mprotect(0x7f5122f6b000, 2097152, PROT_NONE) = 0 <0.000013>
[pid  5087] 11:15:42.169422 mmap(0x7f512316b000, 8192, PROT_READ|PROT_WRITE,
   MAP_PRIVATE|MAP_FIXED|MAP_DENYWRITE, 8, 0x2000) = 0x7f512316b000 <0.000018>
[pid  5087] 11:15:42.169480 close(8)    = 0 <0.000008>
[pid  5087] 11:15:42.169547 mprotect(0x7f512316b000, 4096, PROT_READ) = 0 <0.000014>
[pid  5087] 11:15:42.170460
   open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/site_ruby/2.4.0/enc/trans/transdb.so",
   O_RDONLY|O_NONBLOCK|O_CLOEXEC) = -1 ENOENT (No such file or directory) <0.000015>
[pid  5087] 11:15:42.170521
   open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/site_ruby/2.4.0/x86_64-linux/enc/trans/transdb.so",
   O_RDONLY|O_NONBLOCK|O_CLOEXEC) = -1 ENOENT (No such file or directory) <0.000020>
```

System Calls X

```
[pid  5087] 11:15:42.170586
   open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/site_ruby/enc/trans/transdb.so",
   O_RDONLY|O_NONBLOCK|O_CLOEXEC) = -1 ENOENT (No such file or directory) <0.000012>
[pid  5087] 11:15:42.170636
   open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/vendor_ruby/2.4.0/enc/trans/transdb.so",
   O_RDONLY|O_NONBLOCK|O_CLOEXEC) = -1 ENOENT (No such file or directory) <0.000015>
[pid  5087] 11:15:42.170688
   open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/vendor_ruby/2.4.0/x86_64-linux/enc/trans/transdb.so",
   O_RDONLY|O_NONBLOCK|O_CLOEXEC) = -1 ENOENT (No such file or directory) <0.000012>
[pid  5087] 11:15:42.170740
   open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/vendor_ruby/enc/trans/transdb.so",
   O_RDONLY|O_NONBLOCK|O_CLOEXEC) = -1 ENOENT (No such file or directory) <0.000012>
[pid  5087] 11:15:42.170787 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/2.4.0/enc/trans/transdb.so",
   O_RDONLY|O_NONBLOCK|O_CLOEXEC) = -1 ENOENT (No such file or directory) <0.000012>
[pid  5087] 11:15:42.170833
   open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/2.4.0/x86_64-linux/enc/trans/transdb.so",
   O_RDONLY|O_NONBLOCK|O_CLOEXEC) = 8 <0.000023>
[pid  5087] 11:15:42.170888 fstat(8, {st_mode=S_IFREG|0755, st_size=20200, ...}) = 0 <0.000010>
[pid  5087] 11:15:42.170930 close(8)    = 0 <0.000009>
[pid  5087] 11:15:42.170992
   open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/2.4.0/x86_64-linux/enc/trans/transdb.so",
   O_RDONLY|O_CLOEXEC) = 8 <0.000013>
[pid  5087] 11:15:42.171025 read(8, "\177ELF\2\1\1\0\0\0\0\0\0\0\0\0\3\0>\0\1\0\0\0\0\6\0\0\0\0\0\0"...,
   832) = 832 <0.000009>
.................
.................
[pid  5087] 11:15:42.320297 write(1, "hello world", 11hello world) = 11 <0.000015>
[pid  5087] 11:15:42.320361 write(1, "\n", 1
) = 1 <0.000022>
.................
```

System Calls XI

```
[pid  5106] 11:15:42.322634 exit(0)     = ?
[pid  5087] 11:15:42.322715 <... futex resumed> ) = 0 <0.000299>
[pid  5106] 11:15:42.322764 +++ exited with 0 +++
11:15:42.322787 munmap(0x7f51248ee000, 1052672) = 0 <0.000040>
11:15:42.322863 munmap(0x7f51249f7000, 200704) = 0 <0.000042>
11:15:42.322997 exit_group(0)           = ?
11:15:42.324387 +++ exited with 0 +++

########################################################

# Trace all 'open' syscalls
sudo strace -ttT -ff -e trace=open ruby -e 'puts "hello world"'
11:28:40.909716 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/tls/x86_64/libruby.so.2.4", O_RDONLY|O_CLOEXEC) =
11:28:40.909820 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/tls/libruby.so.2.4", O_RDONLY|O_CLOEXEC) = -1 ENO
11:28:40.909863 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/x86_64/libruby.so.2.4", O_RDONLY|O_CLOEXEC) = -1
11:28:40.909902 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/libruby.so.2.4", O_RDONLY|O_CLOEXEC) = 3 <0.00001
11:28:40.910057 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/libpthread.so.0", O_RDONLY|O_CLOEXEC) = -1 ENOENT
11:28:40.910085 open("/etc/ld.so.cache", O_RDONLY|O_CLOEXEC) = 3 <0.000008>
11:28:40.910155 open("/usr/lib/libpthread.so.0", O_RDONLY|O_CLOEXEC) = 3 <0.000008>
11:28:40.910282 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/libgmp.so.10", O_RDONLY|O_CLOEXEC) = -1 ENOENT (N
11:28:40.910307 open("/usr/lib/libgmp.so.10", O_RDONLY|O_CLOEXEC) = 3 <0.000007>
11:28:40.910418 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/libdl.so.2", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No
11:28:40.910445 open("/usr/lib/libdl.so.2", O_RDONLY|O_CLOEXEC) = 3 <0.000008>
11:28:40.910555 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/libcrypt.so.1", O_RDONLY|O_CLOEXEC) = -1 ENOENT (
11:28:40.910580 open("/usr/lib/libcrypt.so.1", O_RDONLY|O_CLOEXEC) = 3 <0.000008>
11:28:40.910702 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/libm.so.6", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No s
11:28:40.910727 open("/usr/lib/libm.so.6", O_RDONLY|O_CLOEXEC) = 3 <0.000008>
11:28:40.910838 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/libc.so.6", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No s
11:28:40.910863 open("/usr/lib/libc.so.6", O_RDONLY|O_CLOEXEC) = 3 <0.000007>
11:28:40.911636 open("/usr/lib/locale/locale-archive", O_RDONLY|O_CLOEXEC) = 3 <0.000009>
```

System Calls XII

```
11:28:40.911784 open("/proc/self/maps", O_RDONLY|O_CLOEXEC) = 3 <0.000017>
strace: Process 3063 attached
[pid  3062] 11:28:40.921043 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/site_ruby/2.4.0/enc/encdb.so", O
[pid  3062] 11:28:40.921099 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/site_ruby/2.4.0/x86_64-linux/enc
[pid  3062] 11:28:40.921130 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/site_ruby/enc/encdb.so", O_RDONL
[pid  3062] 11:28:40.921160 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/vendor_ruby/2.4.0/enc/encdb.so",
[pid  3062] 11:28:40.921195 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/vendor_ruby/2.4.0/x86_64-linux/e
[pid  3062] 11:28:40.921238 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/vendor_ruby/enc/encdb.so", O_RDO
[pid  3062] 11:28:40.921282 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/2.4.0/enc/encdb.so", O_RDONLY|O_
[pid  3062] 11:28:40.921319 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/2.4.0/x86_64-linux/enc/encdb.so"
[pid  3062] 11:28:40.921471 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/2.4.0/x86_64-linux/enc/encdb.so"
[pid  3062] 11:28:40.922454 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/site_ruby/2.4.0/enc/trans/transd
[pid  3062] 11:28:40.922493 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/site_ruby/2.4.0/x86_64-linux/enc
[pid  3062] 11:28:40.922522 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/site_ruby/enc/trans/transdb.so",
[pid  3062] 11:28:40.922551 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/vendor_ruby/2.4.0/enc/trans/tran
[pid  3062] 11:28:40.922584 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/vendor_ruby/2.4.0/x86_64-linux/e
.................
.................
[pid  3062] 11:28:41.038858 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/gems/2.4.0/gems/did_you_mean-1.1
[pid  3062] 11:28:41.039000 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/gems/2.4.0/gems/did_you_mean-1.1
[pid  3062] 11:28:41.039980 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/gems/2.4.0/gems/did_you_mean-1.1
[pid  3062] 11:28:41.040111 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/gems/2.4.0/gems/did_you_mean-1.1
[pid  3062] 11:28:41.041014 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/gems/2.4.0/gems/did_you_mean-1.1
[pid  3062] 11:28:41.041144 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/gems/2.4.0/gems/did_you_mean-1.1
[pid  3062] 11:28:41.041765 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/gems/2.4.0/gems/did_you_mean-1.1
[pid  3062] 11:28:41.041893 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/gems/2.4.0/gems/did_you_mean-1.1
hello world
[pid  3063] 11:28:41.044118 +++ exited with 0 +++
11:28:41.045734 +++ exited with 0 +++
```

System Calls XIII

```
########################################################

# Trace all file related syscalls
strace -ttT -ff -e trace=file ruby -e 'puts "hello world"'
11:31:25.733312 execve("/home/suresh/.rvm/rubies/ruby-2.4.1/bin/ruby", ["ruby", "-e", "puts \"hello world\""],
11:31:25.733617 access("/etc/ld.so.preload", R_OK) = -1 ENOENT (No such file or directory) <0.000011>
11:31:25.733668 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/tls/x86_64/libruby.so.2.4", O_RDONLY|O_CLOEXEC) =
11:31:25.733712 stat("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/tls/x86_64", 0x7ffdabc866f0) = -1 ENOENT (No such
11:31:25.733758 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/tls/libruby.so.2.4", O_RDONLY|O_CLOEXEC) = -1 ENO
11:31:25.733795 stat("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/tls", 0x7ffdabc866f0) = -1 ENOENT (No such file o
11:31:25.733831 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/x86_64/libruby.so.2.4", O_RDONLY|O_CLOEXEC) = -1
11:31:25.733865 stat("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/x86_64", 0x7ffdabc866f0) = -1 ENOENT (No such fil
11:31:25.733898 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/libruby.so.2.4", O_RDONLY|O_CLOEXEC) = 3 <0.00001
11:31:25.734074 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/libpthread.so.0", O_RDONLY|O_CLOEXEC) = -1 ENOENT
11:31:25.734109 open("/etc/ld.so.cache", O_RDONLY|O_CLOEXEC) = 3 <0.000010>
11:31:25.734198 open("/usr/lib/libpthread.so.0", O_RDONLY|O_CLOEXEC) = 3 <0.000010>
11:31:25.734343 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/libgmp.so.10", O_RDONLY|O_CLOEXEC) = -1 ENOENT (N
11:31:25.734378 open("/usr/lib/libgmp.so.10", O_RDONLY|O_CLOEXEC) = 3 <0.000010>
11:31:25.734512 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/libdl.so.2", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No
11:31:25.734550 open("/usr/lib/libdl.so.2", O_RDONLY|O_CLOEXEC) = 3 <0.000010>
11:31:25.734709 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/libcrypt.so.1", O_RDONLY|O_CLOEXEC) = -1 ENOENT (
11:31:25.734746 open("/usr/lib/libcrypt.so.1", O_RDONLY|O_CLOEXEC) = 3 <0.000010>
11:31:25.734891 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/libm.so.6", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No s
11:31:25.734926 open("/usr/lib/libm.so.6", O_RDONLY|O_CLOEXEC) = 3 <0.000009>
11:31:25.735063 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/libc.so.6", O_RDONLY|O_CLOEXEC) = -1 ENOENT (No s
11:31:25.735098 open("/usr/lib/libc.so.6", O_RDONLY|O_CLOEXEC) = 3 <0.000009>
11:31:25.735888 open("/usr/lib/locale/locale-archive", O_RDONLY|O_CLOEXEC) = 3 <0.000011>
11:31:25.736051 open("/proc/self/maps", O_RDONLY|O_CLOEXEC) = 3 <0.000019>
strace: Process 9532 attached
[pid  9531] 11:31:25.745984 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/site_ruby/2.4.0/enc/encdb.so", O
```

System Calls XIV

```
[pid  9531] 11:31:25.746057 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/site_ruby/2.4.0/x86_64-linux/enc
[pid  9531] 11:31:25.746100 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/site_ruby/enc/encdb.so", O_RDONL
[pid  9531] 11:31:25.746142 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/vendor_ruby/2.4.0/enc/encdb.so",
[pid  9531] 11:31:25.746184 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/vendor_ruby/2.4.0/x86_64-linux/e
[pid  9531] 11:31:25.746225 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/vendor_ruby/enc/encdb.so", O_RDO
[pid  9531] 11:31:25.746267 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/2.4.0/enc/encdb.so", O_RDONLY|O_
[pid  9531] 11:31:25.746321 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/2.4.0/x86_64-linux/enc/encdb.so"
[pid  9531] 11:31:25.746486 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/2.4.0/x86_64-linux/enc/encdb.so"
[pid  9531] 11:31:25.747511 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/site_ruby/2.4.0/enc/trans/transd
[pid  9531] 11:31:25.747556 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/site_ruby/2.4.0/x86_64-linux/enc
[pid  9531] 11:31:25.747596 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/site_ruby/enc/trans/transdb.so",
[pid  9531] 11:31:25.747641 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/vendor_ruby/2.4.0/enc/trans/tran
[pid  9531] 11:31:25.747687 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/vendor_ruby/2.4.0/x86_64-linux/e
...........
...........
[pid  9531] 11:31:25.888752 open("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/gems/2.4.0/gems/did_you_mean-1.1
[pid  9531] 11:31:25.889119 lstat("/home", {st_mode=S_IFDIR|0755, st_size=4096, ...}) = 0 <0.000016>
[pid  9531] 11:31:25.889176 lstat("/home/suresh", {st_mode=S_IFDIR|0700, st_size=4096, ...}) = 0 <0.000011>
[pid  9531] 11:31:25.889213 lstat("/home/suresh/.rvm", {st_mode=S_IFDIR|0755, st_size=4096, ...}) = 0 <0.000011
[pid  9531] 11:31:25.889252 lstat("/home/suresh/.rvm/rubies", {st_mode=S_IFDIR|0755, st_size=4096, ...}) = 0 <0
[pid  9531] 11:31:25.889293 lstat("/home/suresh/.rvm/rubies/ruby-2.4.1", {st_mode=S_IFDIR|0755, st_size=4096, .
[pid  9531] 11:31:25.889341 lstat("/home/suresh/.rvm/rubies/ruby-2.4.1/lib", {st_mode=S_IFDIR|0755, st_size=409
[pid  9531] 11:31:25.889389 lstat("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby", {st_mode=S_IFDIR|0755, st_siz
[pid  9531] 11:31:25.889436 lstat("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/gems", {st_mode=S_IFDIR|0755, s
[pid  9531] 11:31:25.889485 lstat("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/gems/2.4.0", {st_mode=S_IFDIR|0
[pid  9531] 11:31:25.889534 lstat("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/gems/2.4.0/gems", {st_mode=S_IF
[pid  9531] 11:31:25.889589 lstat("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/gems/2.4.0/gems/did_you_mean-1.
[pid  9531] 11:31:25.889637 lstat("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/gems/2.4.0/gems/did_you_mean-1.
[pid  9531] 11:31:25.889690 lstat("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/gems/2.4.0/gems/did_you_mean-1.
[pid  9531] 11:31:25.889739 lstat("/home/suresh/.rvm/rubies/ruby-2.4.1/lib/ruby/gems/2.4.0/gems/did_you_mean-1.
```

System Calls XV

```
hello world
[pid  9532] 11:31:25.892159 +++ exited with 0 +++
11:31:25.893734 +++ exited with 0 +++

########################################################

# Or use 'perf trace'
perf trace ruby -e 'puts "hello world"'
0.024 ( 0.002 ms): ruby/23452 brk(                                                                  ) = 0x1d430
0.042 ( 0.004 ms): ruby/23452 access(filename: 0xf05adc80, mode: R                                  ) = -1 ENOE
0.050 ( 0.004 ms): ruby/23452 open(filename: 0x88f9ce10, flags: CLOEXEC                             ) = -1 ENOE
0.055 ( 0.002 ms): ruby/23452 stat(filename: 0x88f9ce10, statbuf: 0x7ffd88f9cee0                    ) = -1 ENOE
0.058 ( 0.002 ms): ruby/23452 open(filename: 0x88f9ce10, flags: CLOEXEC                             ) = -1 ENOE
0.062 ( 0.002 ms): ruby/23452 stat(filename: 0x88f9ce10, statbuf: 0x7ffd88f9cee0                    ) = -1 ENOE
0.065 ( 0.002 ms): ruby/23452 open(filename: 0x88f9ce10, flags: CLOEXEC                             ) = -1 ENOE
0.069 ( 0.002 ms): ruby/23452 stat(filename: 0x88f9ce10, statbuf: 0x7ffd88f9cee0                    ) = -1 ENOE
0.072 ( 0.004 ms): ruby/23452 open(filename: 0x88f9ce10, flags: CLOEXEC                             ) = 3
0.077 ( 0.002 ms): ruby/23452 read(fd: 3, buf: 0x7ffd88f9d048, count: 832                           ) = 832
0.081 ( 0.002 ms): ruby/23452 fstat(fd: 3</proc/23452/maps>, statbuf: 0x7ffd88f9cee0                ) = 0
.........................
.........................
114.789 ( 0.009 ms): ruby/23452 write(fd: 1</dev/pts/1>, buf: 0x1f38c20, count: 11                    ) = 11
114.802 ( 0.003 ms): ruby/23452 write(fd: 1</dev/pts/1>, buf: 0x7f5cf02a53fc, count: 1                ) = 1
115.474 ( 0.041 ms): ruby/23452 futex(uaddr: 0x7f5cf07ae9d0, val: 23453, uaddr2: 0xca, val3: 140037148305872) =
  8.781 (106.695 ms): ruby-timer-thr/23453  ... [continued]: poll()) = 2
115.519 (18446744073709.512 ms): ruby/23452 munmap(addr: 0x7f5cf0671000, len: 1052672
115.480 ( 0.002 ms): ruby-timer-thr/23453 read(fd: 3, buf: 0x7f5cf0587000, count: 1024
115.484 ( 0.001 ms): ruby-timer-thr/23453 read(fd: 5, buf: 0x7f5cf0587000, count: 1024
115.487 ( 0.002 ms): ruby-timer-thr/23453 close(fd: 3
115.495 ( 0.002 ms): ruby-timer-thr/23453 close(fd: 5
```

System Calls XVI

```
115.502 ( 0.000 ms): ruby-timer-thr/23453 exit(
115.519 ( 0.016 ms): ruby/23452  ... [continued]: munmap()) = 0
115.539 ( 0.015 ms): ruby/23452 munmap(addr: 0x7f5cf077a000, len: 200704                              ) = 0
115.583 ( 0.000 ms): ruby/23452 exit_group(                                                           )
```

[14] http://duartes.org/gustavo/blog/post/system-calls/

[15] http://www.brendangregg.com/blog/2014-05-11/strace-wow-much-syscall.html

[16] http://man7.org/linux/man-pages/man2/ptrace.2.html

System Calls: Mode Switch vs Context Switch I

- Mode Switch: Switching from userspace to kernel (privilege 3 to privilege 0)
- Context Switch: Switching from process to another (== save all registers etc)
- Mode Switch != Context Switch
- A mode switch may / may not lead to a context switch, depending on the syscall/work involved
- Example: Making thousands of syscalls and not running into any context switches

System Calls: Mode Switch vs Context Switch II

```
# Just make sure there are syscalls being made
strace -c -e getpid ruby -e ’for i in 0..10000; Process.pid; end’
% time     seconds  usecs/call     calls    errors syscall
------ ----------- ----------- --------- --------- ----------------
100.00    0.016122           2     10004           getpid
------ ----------- ----------- --------- --------- ----------------
100.00    0.016122                 10004           total


# Trace the number of context switches
sudo perf stat -e context-switches ruby -e ’for i in 0..10000; Process.pid; end’

 Performance counter stats for ’ruby -e for i in 0..10000; Process.pid; end’:

                 2      context-switches

       0.097093505 seconds time elapsed
```

- Example: Make a http GET, observe how many context switches happen

System Calls: Mode Switch vs Context Switch III

```
# How many context-switches we run into during a simple http GET call
sudo perf stat -e context-switches ruby -e "require ’net/http’; \
    Net::HTTP.get(URI.parse(’https://www.google.co.in/’));"

 Performance counter stats for ’ruby -e require ’net/http’; Net::HTTP.get(URI.parse(’https://www.google.co.in/’

                17      context-switches

       0.336323461 seconds time elapsed


# Check where we are getting scheduled out
sudo perf record --call-graph dwarf -e context-switches \
 ruby -e "require ’net/http’; Net::HTTP.get(URI.parse(’http://google.com/’));"

# Display the callstacks where it was scheduled out
sudo perf report --call-graph=flat,count --stdio  --no-children
```

Various process metrics

- CPU utilization
  - Remember: This also includes waiting to access memory
- Memory utilization
  - Heap
  - Stack
- Disk utilization

Case study: Simple case - Curl I

```
$ strace -ff -ttT -e trace=network curl --silent -o /dev/null http://www.google.co.in
.......
[pid 24286] 18:40:54.282834 socket(AF_INET, SOCK_DGRAM|SOCK_NONBLOCK, IPPROTO_IP) = 3 <0.000020>
[pid 24286] 18:40:54.282889 connect(3, {sa_family=AF_INET, sin_port=htons(53), sin_addr=inet_addr("8.8.8.8")},
.......
[pid 24286] 18:40:54.322050 socket(AF_INET, SOCK_DGRAM, IPPROTO_IP) = 3 <0.000015>
[pid 24286] 18:40:54.322088 connect(3, {sa_family=AF_INET, sin_port=htons(80), sin_addr=inet_addr("172.217.26.1
[pid 24286] 18:40:54.322168 getsockname(3, {sa_family=AF_INET, sin_port=htons(38303), sin_addr=inet_addr("192.1
[pid 24286] 18:40:54.322218 socket(AF_INET6, SOCK_DGRAM, IPPROTO_IP) = 3 <0.000014>
[pid 24286] 18:40:54.322252 connect(3, {sa_family=AF_INET6, sin6_port=htons(80), inet_pton(AF_INET6, "2404:6800
[pid 24286] 18:40:54.322451 +++ exited with 0 +++
18:40:54.340534 socket(AF_INET, SOCK_STREAM, IPPROTO_TCP) = 3 <0.000052>
.......
18:40:54.340881 connect(3, {sa_family=AF_INET, sin_port=htons(80), sin_addr=inet_addr("172.217.26.163")}, 16) =
.......
18:40:54.344268 sendto(3, "GET / HTTP/1.1\r\nHost: www.google"..., 80, MSG_NOSIGNAL, NULL, 0) = 80 <0.000022>
18:40:54.474244 recvfrom(3, "HTTP/1.1 200 OK\r\nDate: Tue, 08 A"..., 102400, 0, NULL, NULL) = 13880 <0.000012>
18:40:54.477151 recvfrom(3, "umpException=window._._DumpExcep"..., 102400, 0, NULL, NULL) = 1275 <0.000008>
18:40:54.478162 +++ exited with 0 +++
```

Case study: How Ruby threading works I

- MRI Ruby is single threaded
- Even though it supports multi-threading, only one thread can execute ruby code at a given time.
- Why: MRI Ruby VM is not multi-thread safe: A global lock must be acquired by a thread before it can execute ruby code. It is called GVL (Global VM Lock) in Ruby (like GIL in python).
- So in a Ruby process, at any given time, only one thread can execute for all practical purposes, everybody else must wait for the currently running thread to release the GVL lock.
- Now there are two scenarios w.r.t multi-threading
    1. Current thread is voluntarily giving up the lock (by calling blocking functions, like sleep, or waiting on network call etc).
    2. Current thread doesn't call any blocking function, just busily does some work (like big regex match, or "loop do end" etc).

Case study: How Ruby threading works II

For case 1, this is the most straightforward, RubyVM doesn't have to do anything. Here is an example, sleep function [17], notice that:

- It releases the GVL lock
- Goes to sleep
- As soon as it wakes up, it will try to acquire the GVL lock, only then it can proceed further. If it can't acquire GVL lock immediately, it will have to wait. For case 2, we will see below how it works.

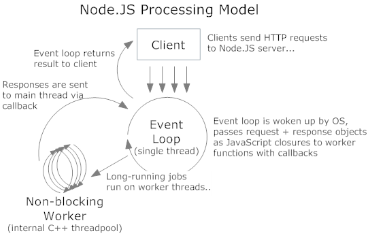
- How Ruby scheduler works:
  - If a thread acquires GVL, if it continues to run without calling any blocking function, it can unfairly starve all other threads (all those threads will be waiting to acquire GVL, which this thread doesn't release).
  - So Ruby implements a sort of co-operative scheduler: a pseudo co-operative interrupt mechanism

Case study: How Ruby threading works III

- A timer thread is started, it keeps running forever. Periodically, it will set a "flag" in the current running thread to indicate it should stop running and give up the GVL. The function that actually does it: `https://github.com/ruby/ruby/blob/ruby_2_2/thread.c#L3817-L3834`
- Now the current thread that is running will check this flag periodically to see if it needs to stop running. If so, it will give up the lock and go back to wait list.

[17] `https://github.com/ruby/ruby/blob/ruby_2_2/thread_pthread.c#L1115-L1136`

Case study: How NodeJS eventing/threading works I



Node.JS Processing Model

Client

Clients send HTTP requests to Node.JS server...

Event loop returns result to client

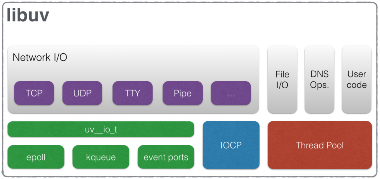
Responses are sent to main thread via callback

Event Loop (single thread)

Event loop is woken up by OS, passes request + response objects as JavaScript closures to worker functions with callbacks

Long-running jobs run on worker threads..

Non-blocking Worker (internal C++ threadpool)



libuv

Network I/O

TCP UDP TTY Pipe ...

File I/O

DNS Ops.

User code

uv__io_t

epoll kqueue event ports

IOCP

Thread Pool

- Node.js has single thread + event loop architecture
- It uses libuv for "handling" events
- How does it handle blocking syscalls?
  - It uses non-blocking IO whereever possible (if OS supports) and fallbacks to threadpool for blocking IO
    - Non-blocking: TCP, UDP, Pipes etc
    - Blocking: DNS lookups, Disk read/writes etc
- Let's analyze it from the PoV of OS to validate this

Case study: How NodeJS eventing/threading works II

```
# strace -ff -ttT -e trace=network,write,read node -p ’console.log(‘##### Process ID ####: ${process.pid}‘); \
   req=http.get("http://www.google.co.in", function(res) { var body=""; \
   res.on("data",function(data){body+=data;}); \
   res.on("end", function() {console.log("======================")} )});console.log("Done")’
.......
[pid 12551] 15:42:26.773798 write(9, "########## Process ID ##########"..., 41########## Process ID ###########
.......
[pid 12551] 15:42:26.799280 write(9, "Done\n", 5) = 5 <0.000018>
.......
[pid 12557] 15:42:26.799574 socket(AF_INET, SOCK_DGRAM|SOCK_NONBLOCK, IPPROTO_IP) = 12 <0.000014>
[pid 12557] 15:42:26.799615 connect(12, {sa_family=AF_INET, sin_port=htons(53), sin_addr=inet_addr("8.8.8.8")},
[pid 12557] 15:42:26.799675 sendmmsg(12, [{msg_hdr={msg_name=NULL, msg_namelen=0, msg_iov=[{iov_base="\377F\1\0
[pid 12557] 15:42:26.838085 recvfrom(12, "\377F\201\200\0\1\0\1\0\0\0\0\3www\6google\2co\2in\0\0\1"..., 2048, 0
.......
[pid 12557] 15:42:26.838542 socket(AF_INET, SOCK_DGRAM, IPPROTO_IP) = 12 <0.000014>
[pid 12557] 15:42:26.838577 connect(12, {sa_family=AF_INET, sin_port=htons(0), sin_addr=inet_addr("216.58.196.9
[pid 12557] 15:42:26.838618 getsockname(12, {sa_family=AF_INET, sin_port=htons(56665), sin_addr=inet_addr("192.
[pid 12557] 15:42:26.838666 socket(AF_INET6, SOCK_DGRAM, IPPROTO_IP) = 12 <0.000013>
[pid 12557] 15:42:26.838713 connect(12, {sa_family=AF_INET6, sin6_port=htons(0), inet_pton(AF_INET6, "2404:6800
.......
[pid 12551] 15:42:26.839723 socket(AF_INET, SOCK_STREAM|SOCK_CLOEXEC|SOCK_NONBLOCK, IPPROTO_IP) = 12 <0.000052>
[pid 12551] 15:42:26.839811 connect(12, {sa_family=AF_INET, sin_port=htons(80), sin_addr=inet_addr("216.58.196.
[pid 12551] 15:42:26.842964 getsockopt(12, SOL_SOCKET, SO_ERROR, [0], [4]) = 0 <0.000010>
[pid 12551] 15:42:26.843780 write(12, "GET / HTTP/1.1\r\nHost: www.google"..., 61) = 61 <0.000038>
[pid 12551] 15:42:26.968154 read(12, "HTTP/1.1 200 OK\r\nDate: Mon, 07 A"..., 65536) = 13880 <0.000024>
[pid 12551] 15:42:26.972625 read(12, "f1f1);background-image:-ms-linea"..., 65536) = 27760 <0.000025>
[pid 12551] 15:42:26.974256 read(12, "<a class=gbmt href=\"http://www.g"..., 65536) = 7844 <0.000016>
[pid 12551] 15:42:26.974416 read(12, "", 65536) = 0 <0.000011>
[pid 12551] 15:42:26.976656 write(9, "======================\n", 23) = 23 <0.000027>
.......
```

Case study: How NodeJS eventing/threading works III

- **Observations**
  - We do see the DNS look up being made from a separate thread
  - Making http request(= `TCP)`, `writing to console(=` TTY) etc are being made from main thread
  - But we also see multiple 'connect' calls to the server, why and where do they originate from?

```
sudo perf trace -T -e connect --call-graph=dwarf node -p 'console.log(`##### Process ID ####:${process.pid}`);\
   req=http.get("http://www.google.co.in", function(res) { var body=""; \
   res.on("data",function(data){body+=data;}); \
   res.on("end", function() {console.log("===================")} )});console.log("Done")'
########## Process ID ###########: 10113
Done
160498846.387 ( 0.026 ms): :10119/10119 connect(fd: 13, uservaddr: 0x7fbce3e345a0, addrlen: 110            )
                                        [0] ([unknown])
160498846.432 ( 0.004 ms): :10119/10119 connect(fd: 12<socket:[32719970]>, uservaddr: 0x7fbce3e35660, addrlen:
                                        [0] ([unknown])
160498847.194 ( 0.016 ms): :10119/10119 connect(fd: 12<socket:[32719970]>, uservaddr: 0x7fbcdc002e08, addrlen:
                                        [0] ([unknown])
160498848.516 ( 0.014 ms): node/10119 connect(fd: 12<socket:[32719970]>, uservaddr: 0x7fbce3e36dcc, addrlen: 16
                                        __GI___libc_connect (/usr/lib/libpthread-2.25.so)
                                        [0xffff80431ec83d32] (/usr/lib/libresolv-2.25.so)
                                        [0xffff80431ec84c1e] (/usr/lib/libresolv-2.25.so)
                                        __libc_res_nquery (/usr/lib/libresolv-2.25.so)
                                        [0xffff80431ec833af] (/usr/lib/libresolv-2.25.so)
                                        __libc_res_nsearch (/usr/lib/libresolv-2.25.so)
```

Case study: How NodeJS eventing/threading works IV

```
                                       _nss_dns_gethostbyname4_r (/usr/lib/libnss_dns-2.25.so)
                                       gaih_inet.constprop.5 (/usr/lib/libc-2.25.so)
                                       getaddrinfo (/usr/lib/libc-2.25.so)
                                       [0xffff8043161249e9] (/usr/lib/libuv.so.1.0.0)
                                       [0xffff80431611ca54] (/usr/lib/libuv.so.1.0.0)
                                       start_thread (/usr/lib/libpthread-2.25.so)
                                       __clone (/usr/lib/libc-2.25.so)
160498887.935 ( 0.014 ms): node/10119 connect(fd: 12<socket:[32719970]>, uservaddr: 0x7fbcdc001fb0, addrlen: 16
                                       __GI___libc_connect (/usr/lib/libc-2.25.so)
                                       getaddrinfo (/usr/lib/libc-2.25.so)
160498887.964 ( 0.018 ms): node/10119 connect(fd: 12<socket:[32719970]>, uservaddr: 0x7fbcdc002000, addrlen: 28
                                       __GI___libc_connect (/usr/lib/libc-2.25.so)
                                       getaddrinfo (/usr/lib/libc-2.25.so)
Failed to open /tmp/perf-10113.map, continuing without symbols
160498889.020 ( 0.058 ms): node/10113 connect(fd: 12<socket:[32718959]>, uservaddr: 0x7ffebcc03020, addrlen: 16
                                       __GI___libc_connect (/usr/lib/libpthread-2.25.so)
                                       uv__tcp_connect (/usr/lib/libuv.so.1.0.0)
                                       node::TCPWrap::Connect (/usr/bin/node)
                                       v8::internal::FunctionCallbackArguments::Call (/usr/bin/node)
                                       [0xf0b8] (/usr/bin/node)
                                       [0xf450] (/usr/bin/node)
                                       v8::internal::Builtin_HandleApiCall (/usr/bin/node)
                                       [0x3dbac15043a7] (/tmp/perf-10113.map)
```

Case study: How Golang threading works I

TODO

Topic

1. System

2. Application

3. Profilers

4. Debuggers

5. Books

Profilers

- Sampling vs Precise
  - Tail latency
- Use cases
  - CPU intensive
    - CPU cache usage and Cache contention
  - Blocking (off-cpu)
  - Language specific
- perf
- stackprof

Perf I

- Core vs HyperThreading https://github.com/andikleen/pmu-tools https://github.com/andikleen/pmu-tools/wiki/toplev-manual

```
# Run it on the same core (HyperThreading)
echo -n "2 6" | xargs -I’{}’ --delimit ’ ’ --max-args=1 --max-procs=$(nproc) \
  sudo sh -c ’chrt -f 99 dd if=/dev/zero bs=8M count=128 | \
  toplev.py --quiet --single-thread -l3 taskset -c {} chrt -f 99 gzip > /dev/null’
128+0 records in
128+0 records out
1073741824 bytes (1.1 GB, 1.0 GiB) copied, 14.2675 s, 75.3 MB/s
BE             Backend_Bound:                                60.77 % Slots  [  6.25%]
BE/Mem         Backend_Bound.Memory_Bound:                   44.44 % Slots  [  6.25%]
BE/Mem         Backend_Bound.Memory_Bound.L1_Bound:          25.39 % Clocks [  6.26%] BN

128+0 records in
128+0 records out
1073741824 bytes (1.1 GB, 1.0 GiB) copied, 14.3168 s, 75.0 MB/s
BE             Backend_Bound:                                51.97 % Slots  [  6.25%]
BE/Mem         Backend_Bound.Memory_Bound:                   43.18 % Slots  [  6.25%]
BE/Mem         Backend_Bound.Memory_Bound.L1_Bound:          46.23 % Clocks [  6.26%] BN


# Run it on different cores
echo -n "2 3" | xargs -I’{}’ --delimit ’ ’ --max-args=1 --max-procs=$(nproc) \
  sudo sh -c ’chrt -f 99 dd if=/dev/zero bs=8M count=128 | \
```

Perf II

```
  toplev.py --quiet --single-thread -l3 taskset -c {} chrt -f 99 gzip > /dev/null'

128+0 records in
128+0 records out
1073741824 bytes (1.1 GB, 1.0 GiB) copied, 9.90494 s, 108 MB/s
128+0 records in
128+0 records out
1073741824 bytes (1.1 GB, 1.0 GiB) copied, 9.90628 s, 108 MB/s
BE             Backend_Bound:                                      42.88 % Slots  [  6.29%] BN
BE/Mem         Backend_Bound.Memory_Bound:                         21.33 % Slots  [  6.22%]
BE/Core        Backend_Bound.Core_Bound:                           21.55 % Slots  [  6.22%]
BE/Mem         Backend_Bound.Memory_Bound.L1_Bound:                32.82 % Clocks [  6.20%]
BE/Core        Backend_Bound.Core_Bound.Ports_Utilization:         36.13 % Clocks [  6.26%]
BE             Backend_Bound:                                      42.66 % Slots  [  6.25%] BN
BE/Mem         Backend_Bound.Memory_Bound:                         20.95 % Slots  [  6.25%]
BE/Core        Backend_Bound.Core_Bound:                           21.71 % Slots  [  6.25%]
BE/Mem         Backend_Bound.Memory_Bound.L1_Bound:                32.87 % Clocks [  6.22%]
BE/Core        Backend_Bound.Core_Bound.Ports_Utilization:         36.66 % Clocks [  6.18%]
```

Perf III

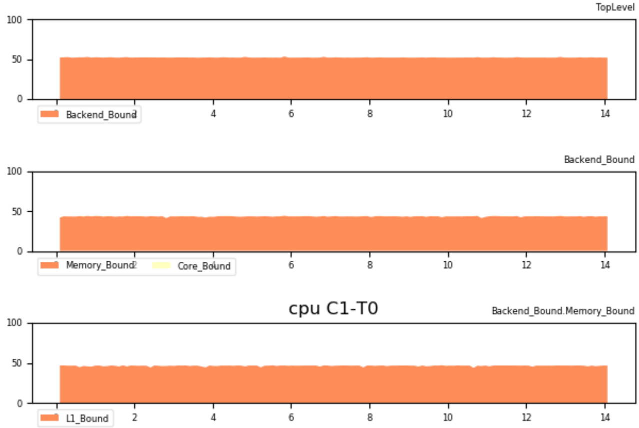


Figure: Performance on shared core

Perf IV

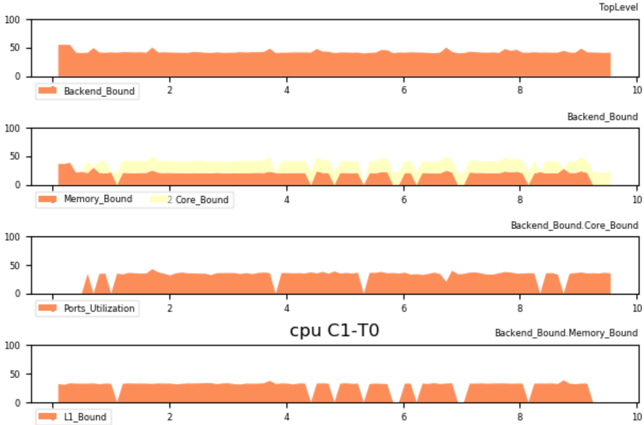


Figure: Performance on dedicated core

Topic

1. System
2. Application
3. Profilers
4. Debuggers
5. Books

Debuggers

- gdb
- rbtrace

Topic

1. System
2. Application
3. Profilers
4. Debuggers
5. Books

Books for reference I

- How computers work
  - Code: The Hidden Language of Computer Hardware and Software
  - Hardware
    - The Indispensable PC Hardware Book Hardcover by Mr Hans-Peter Messmer
    - Modern Processor Design: Fundamentals of Superscalar Processors by John Paul Shen
    - Pentium Pro and Pentium II System Architecture (2nd Edition) by Tom Shanley
  - OS
    - Operating Systems: Three Easy Pieces - `http://pages.cs.wisc.edu/~remzi/OSTEP/`
    - xv6 - `https://pdos.csail.mit.edu/6.828/2014/xv6/book-rev8.pdf`
- Linux
  - Linux Kernel Development by Robert Love

Books for reference II

- The Linux Programming Interface – A Linux and UNIX System Programming Handbook by Michael Kerrisk
- Linux System Programming (2 edition) by Robert Love
- Understanding The Linux Network Internals by Benvenuti

- Performance related
  - Site Reliability Engineering: How Google Runs Production Systems by Niall Murphy, Jennifer Petoff, Chris Jones
  - The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation and Modeling by Raj Jain
  - Systems Performance: Enterprise and the Cloud by Brendan Gregg