DATASTAX

Diagnosing Problems in Production

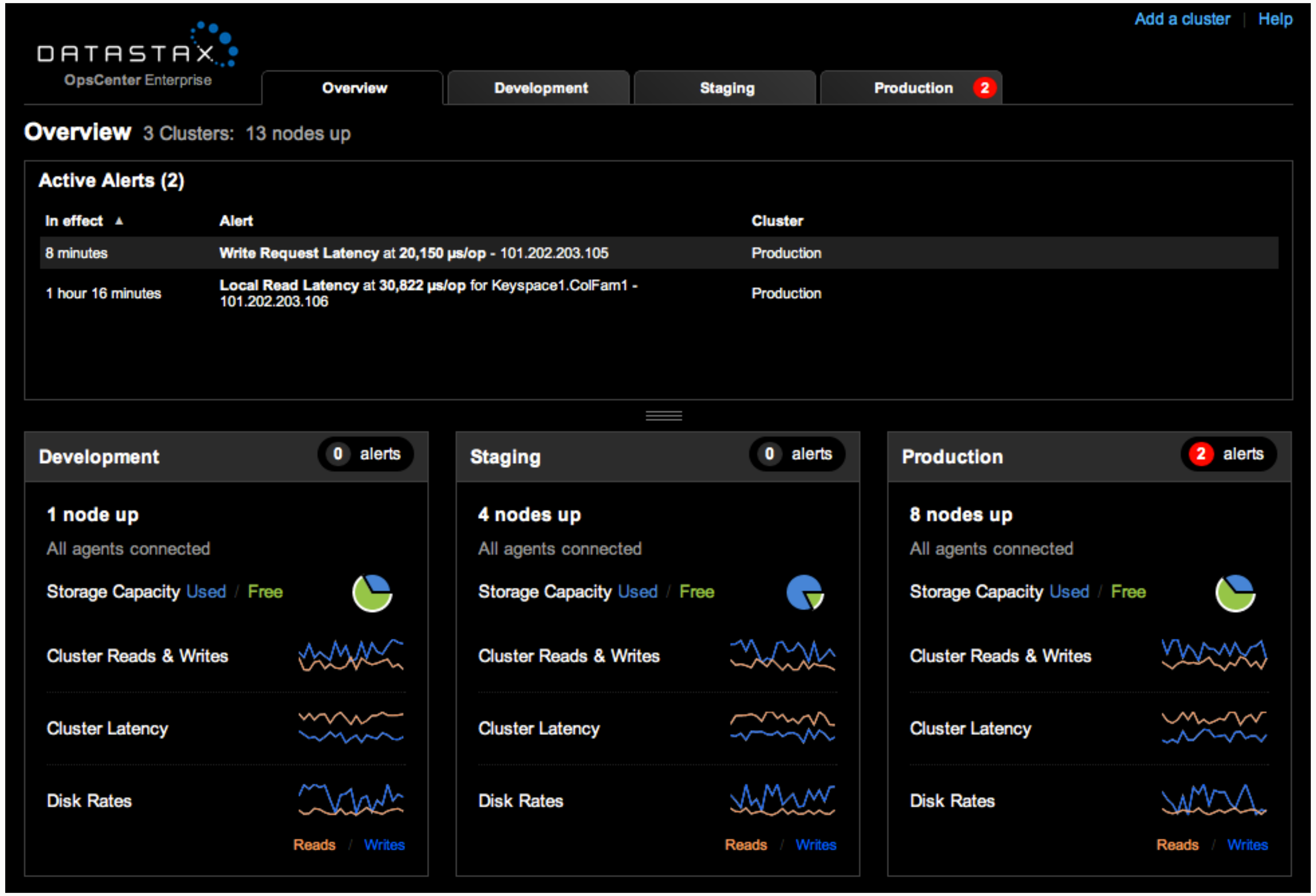
Jon Haddad, Technical Evangelist

@rustyrazorblade

First Step: Preparation

DataStax OpsCenter

- Will help with 90% of problems you encounter
- Should be first place you look when there's an issue
- Community version is free
- Enterprise version has additional features



DATASTAX OpsCenter Enterprise

Add a cluster | Help

Overview | Development | Staging | Production 2

Overview 3 Clusters: 13 nodes up

Active Alerts (2)

In effect ▲	Alert	Cluster
8 minutes	**Write Request Latency** at 20,150 µs/op - 101.202.203.105	Production
1 hour 16 minutes	**Local Read Latency** at 30,822 µs/op for Keyspace1.ColFam1 - 101.202.203.106	Production

Development 0 alerts

1 node up

All agents connected

Storage Capacity Used / Free

Cluster Reads & Writes

Cluster Latency

Disk Rates

Reads / Writes

Staging 0 alerts

4 nodes up

All agents connected

Storage Capacity Used / Free

Cluster Reads & Writes

Cluster Latency

Disk Rates

Reads / Writes

Production 2 alerts

8 nodes up

All agents connected

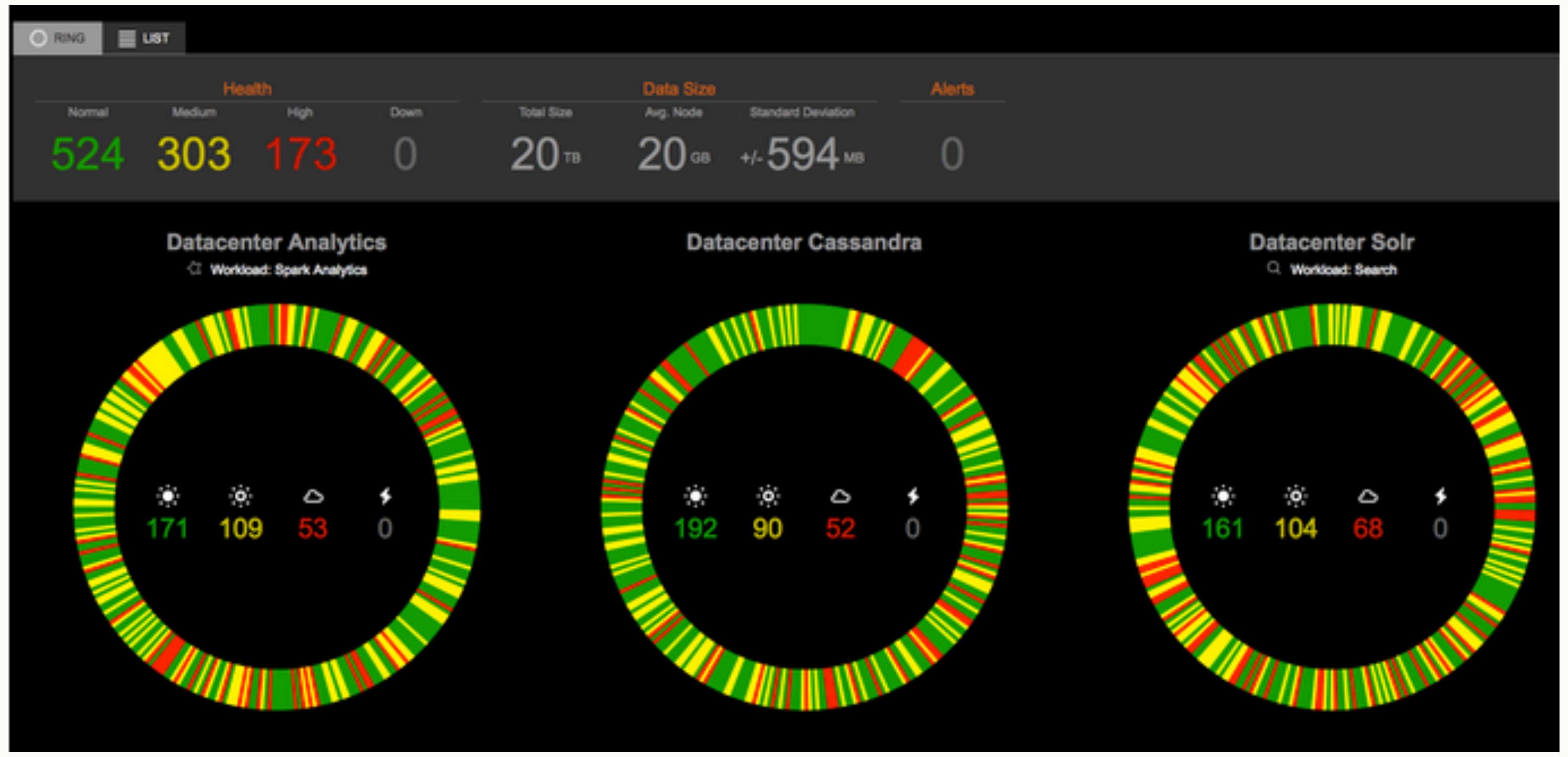
Storage Capacity Used / Free

Cluster Reads & Writes

Cluster Latency

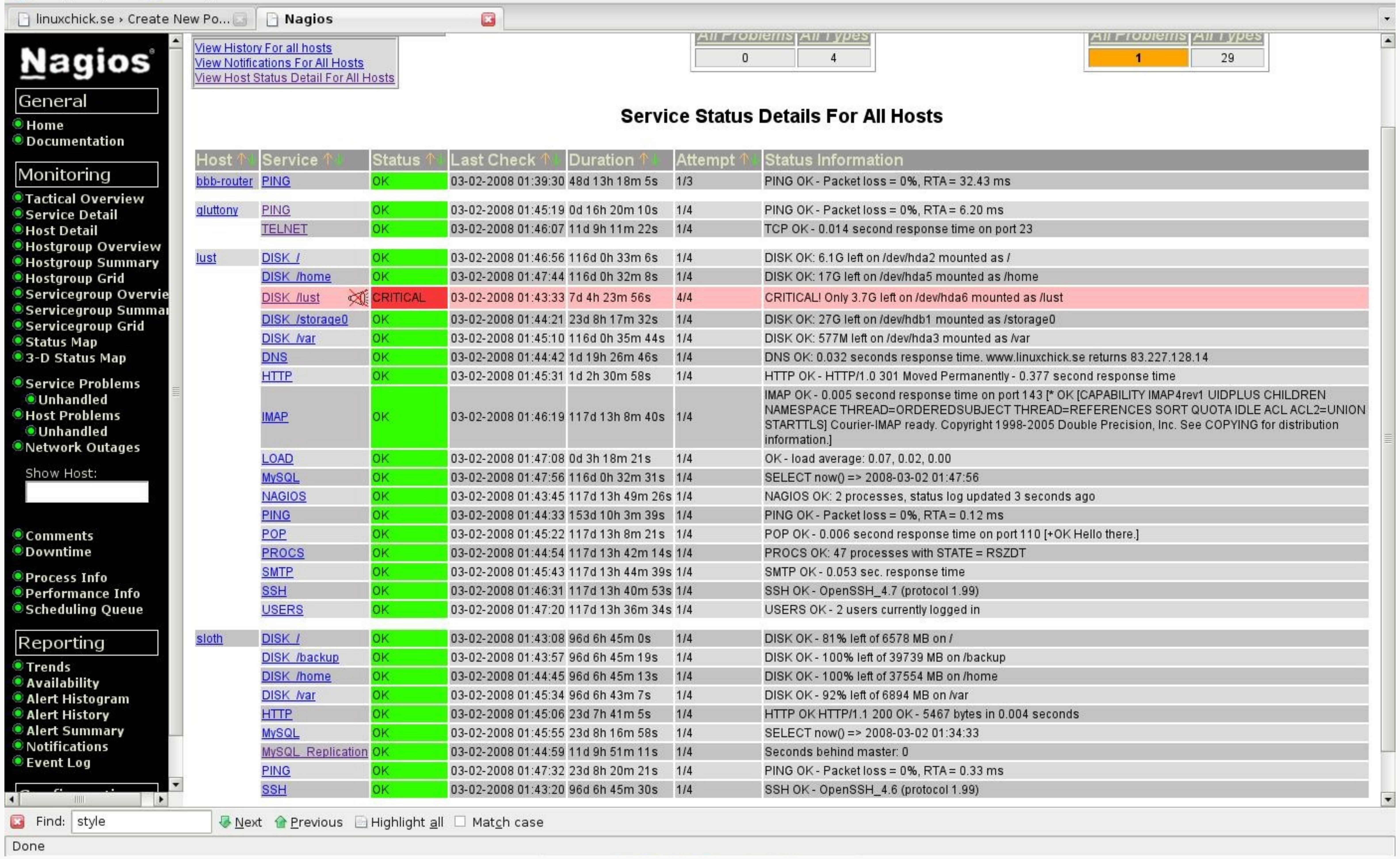
Disk Rates

Reads / Writes



Server Monitoring & Alerts

- Monit
  - monitor processes
  - monitor disk usage
  - send alerts
- Munin / collectd
  - system perf statistics
- Nagios / Icinga
- Various 3rd party services
- Use whatever works for you



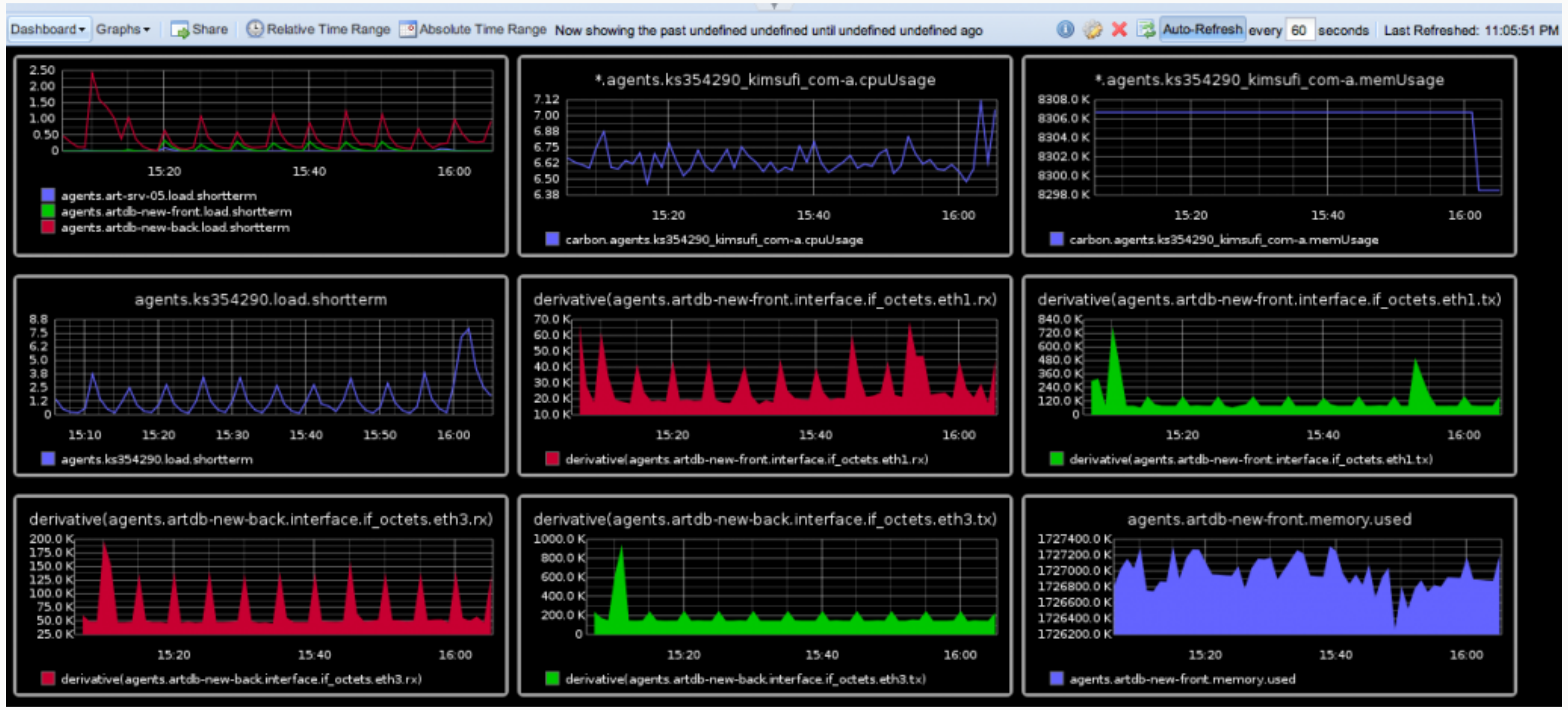
View History For all hosts
View Notifications For All Hosts
View Host Status Detail For All Hosts

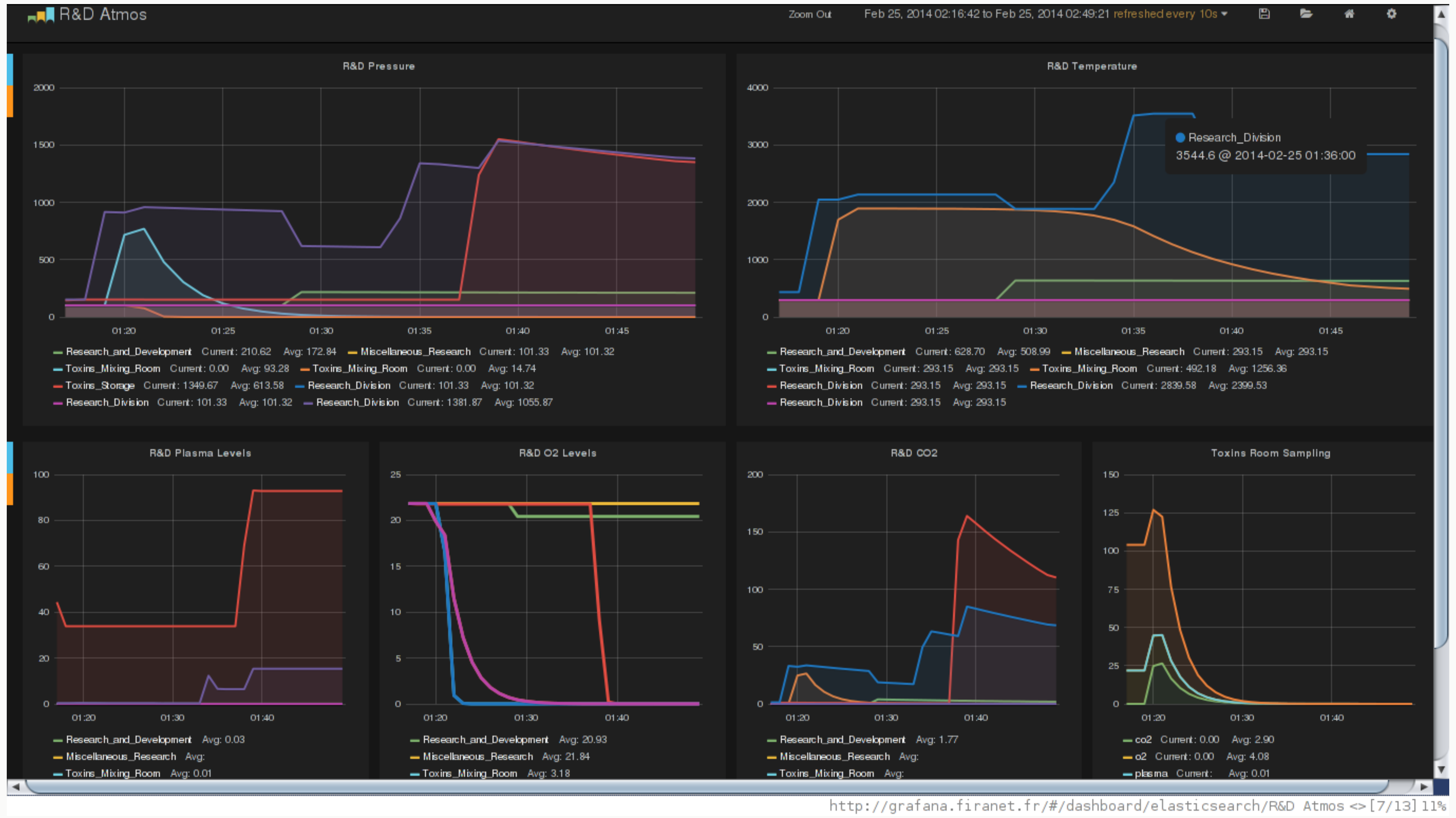
Service Status Details For All Hosts

Host	Service	Status	Last Check	Duration	Attempt	Status Information
bbb-router	PING	OK	03-02-2008 01:39:30	48d 13h 18m 5s	1/3	PING OK - Packet loss = 0%, RTA = 32.43 ms
gluttony	PING	OK	03-02-2008 01:45:19	0d 16h 20m 10s	1/4	PING OK - Packet loss = 0%, RTA = 6.20 ms
	TELNET	OK	03-02-2008 01:46:07	11d 9h 11m 22s	1/4	TCP OK - 0.014 second response time on port 23
lust	DISK /	OK	03-02-2008 01:46:56	116d 0h 33m 6s	1/4	DISK OK: 6.1G left on /dev/hda2 mounted as /
	DISK /home	OK	03-02-2008 01:47:44	116d 0h 32m 8s	1/4	DISK OK: 17G left on /dev/hda5 mounted as /home
	DISK /lust	CRITICAL	03-02-2008 01:43:33	7d 4h 23m 56s	4/4	CRITICAL! Only 3.7G left on /dev/hda6 mounted as /lust
	DISK /storage0	OK	03-02-2008 01:44:21	23d 8h 17m 32s	1/4	DISK OK: 27G left on /dev/hdb1 mounted as /storage0
	DISK /var	OK	03-02-2008 01:45:10	116d 0h 35m 44s	1/4	DISK OK: 577M left on /dev/hda3 mounted as /var
	DNS	OK	03-02-2008 01:44:42	1d 19h 26m 46s	1/4	DNS OK: 0.032 seconds response time. www.linuxchick.se returns 83.227.128.14
	HTTP	OK	03-02-2008 01:45:31	1d 2h 30m 58s	1/4	HTTP OK - HTTP/1.0 301 Moved Permanently - 0.377 second response time
	IMAP	OK	03-02-2008 01:46:19	117d 13h 8m 40s	1/4	IMAP OK - 0.005 second response time on port 143 [* OK [CAPABILITY IMAP4rev1 UIDPLUS CHILDREN NAMESPACE THREAD=ORDEREDSUBJECT THREAD=REFERENCES SORT QUOTA IDLE ACL ACL2=UNION STARTTLS] Courier-IMAP ready. Copyright 1998-2005 Double Precision, Inc. See COPYING for distribution information.]
	LOAD	OK	03-02-2008 01:47:08	0d 3h 18m 21s	1/4	OK - load average: 0.07, 0.02, 0.00
	MySQL	OK	03-02-2008 01:47:56	116d 0h 32m 31s	1/4	SELECT now() => 2008-03-02 01:47:56
	NAGIOS	OK	03-02-2008 01:43:45	117d 13h 49m 26s	1/4	NAGIOS OK: 2 processes, status log updated 3 seconds ago
	PING	OK	03-02-2008 01:44:33	153d 10h 3m 39s	1/4	PING OK - Packet loss = 0%, RTA = 0.12 ms
	POP	OK	03-02-2008 01:45:22	117d 13h 8m 21s	1/4	POP OK - 0.006 second response time on port 110 [+OK Hello there.]
	PROCS	OK	03-02-2008 01:44:54	117d 13h 42m 14s	1/4	PROCS OK: 47 processes with STATE = RSZDT
	SMTP	OK	03-02-2008 01:45:43	117d 13h 44m 39s	1/4	SMTP OK - 0.053 sec. response time
	SSH	OK	03-02-2008 01:46:31	117d 13h 40m 53s	1/4	SSH OK - OpenSSH_4.7 (protocol 1.99)
	USERS	OK	03-02-2008 01:47:20	117d 13h 36m 34s	1/4	USERS OK - 2 users currently logged in
sloth	DISK /	OK	03-02-2008 01:43:08	96d 6h 45m 0s	1/4	DISK OK - 81% left of 6578 MB on /
	DISK /backup	OK	03-02-2008 01:43:57	96d 6h 45m 19s	1/4	DISK OK - 100% left of 39739 MB on /backup
	DISK /home	OK	03-02-2008 01:44:45	96d 6h 45m 13s	1/4	DISK OK - 100% left of 37554 MB on /home
	DISK /var	OK	03-02-2008 01:45:34	96d 6h 43m 7s	1/4	DISK OK - 92% left of 6894 MB on /var
	HTTP	OK	03-02-2008 01:45:06	23d 7h 41m 5s	1/4	HTTP OK HTTP/1.1 200 OK - 5467 bytes in 0.004 seconds
	MySQL	OK	03-02-2008 01:45:55	23d 8h 16m 58s	1/4	SELECT now() => 2008-03-02 01:34:33
	MySQL Replication	OK	03-02-2008 01:44:59	11d 9h 51m 11s	1/4	Seconds behind master: 0
	PING	OK	03-02-2008 01:47:32	23d 8h 20m 21s	1/4	PING OK - Packet loss = 0%, RTA = 0.33 ms
	SSH	OK	03-02-2008 01:43:20	96d 6h 45m 30s	1/4	SSH OK - OpenSSH_4.6 (protocol 1.99)

Application Metrics

- Statsd / Graphite
- Grafana
- Gather constant metrics from your application
- Measure anything & everything
- Microtimers, counters
- Graph events
  - user signup
  - error rates
- Cassandra Metrics Integration
- jmxtrans

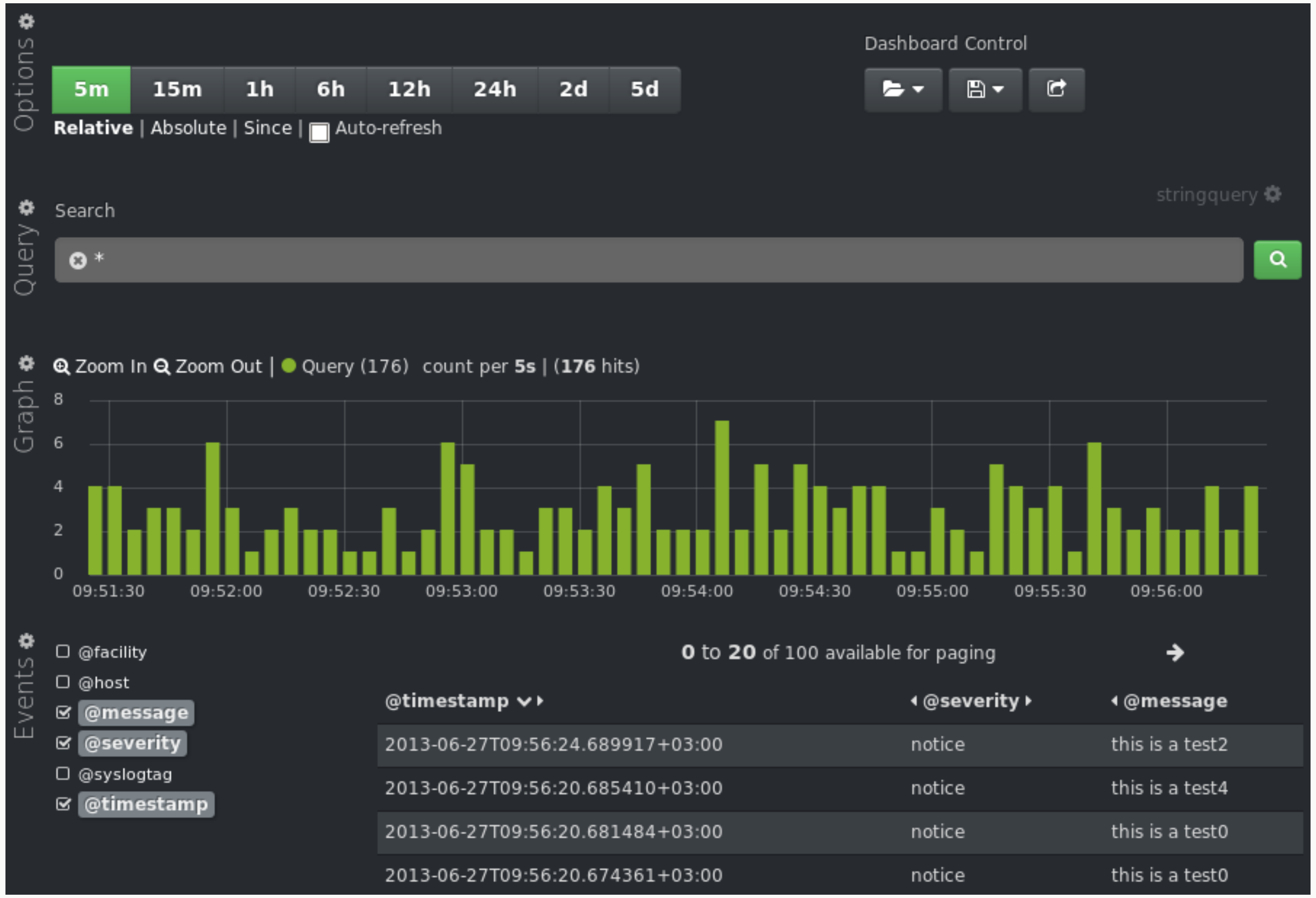

Dashboard ▾
Graphs ▾
Share
Relative Time Range
Absolute Time Range
Now showing the past undefined undefined until undefined undefined ago
Auto-Refresh every 60 seconds
Last Refreshed: 11:05:51 PM
agents.art-srv-05.load.shortterm
agents.artdb-new-front.load.shortterm
agents.artdb-new-back.load.shortterm
*.agents.ks354290_kimsufi_com-a.cpuUsage
carbon.agents.ks354290_kimsufi_com-a.cpuUsage
*.agents.ks354290_kimsufi_com-a.memUsage
carbon.agents.ks354290_kimsufi_com-a.memUsage
agents.ks354290.load.shortterm
derivative(agents.artdb-new-front.interface.if_octets.eth1.rx)
derivative(agents.artdb-new-front.interface.if_octets.eth1.tx)
derivative(agents.artdb-new-back.interface.if_octets.eth3.rx)
derivative(agents.artdb-new-back.interface.if_octets.eth3.tx)
agents.artdb-new-front.memory.used
15:20
15:40
16:00


R&D Atmos
Zoom Out
Feb 25, 2014 02:16:42 to Feb 25, 2014 02:49:21 refreshed every 10s ▾
R&D Pressure
Research_and_Development Current: 210.62 Avg: 172.84
Miscellaneous_Research Current: 101.33 Avg: 101.32
Toxins_Mixing_Room Current: 0.00 Avg: 93.28
Toxins_Mixing_Room Current: 0.00 Avg: 14.74
Toxins_Storage Current: 1349.67 Avg: 613.58
Research_Division Current: 101.33 Avg: 101.32
Research_Division Current: 101.33 Avg: 101.32
Research_Division Current: 1381.87 Avg: 1055.87
R&D Temperature
Research_Division
3544.6 @ 2014-02-25 01:36:00
Research_and_Development Current: 628.70 Avg: 508.99
Miscellaneous_Research Current: 293.15 Avg: 293.15
Toxins_Mixing_Room Current: 293.15 Avg: 293.15
Toxins_Mixing_Room Current: 492.18 Avg: 1256.36
Research_Division Current: 293.15 Avg: 293.15
Research_Division Current: 2839.58 Avg: 2399.53
Research_Division Current: 293.15 Avg: 293.15
R&D Plasma Levels
Research_and_Development Avg: 0.03
Miscellaneous_Research Avg:
Toxins_Mixing_Room Avg: 0.01
R&D O2 Levels
Research_and_Development Avg: 20.93
Miscellaneous_Research Avg: 21.84
Toxins_Mixing_Room Avg: 3.18
R&D CO2
Research_and_Development Avg: 1.77
Miscellaneous_Research Avg:
Toxins_Mixing_Room Avg:
Toxins Room Sampling
co2 Current: 0.00 Avg: 2.90
o2 Current: 0.00 Avg: 4.08
plasma Current: Avg: 0.01
01:20
01:30
01:40

http://grafana.firanet.fr/#/dashboard/elasticsearch/R&D Atmos <>[7/13] 11%

Log Aggregation

- Hosted - Splunk, Loggly
- OSS - Logstash + Kibana, Greylog
- Many more…
- For best results all logs should be aggregated here
- Oh yeah, and log your errors.



- □ @facility
- □ @host
- ☑ @message
- ☑ @severity
- □ @syslogtag
- ☑ @timestamp

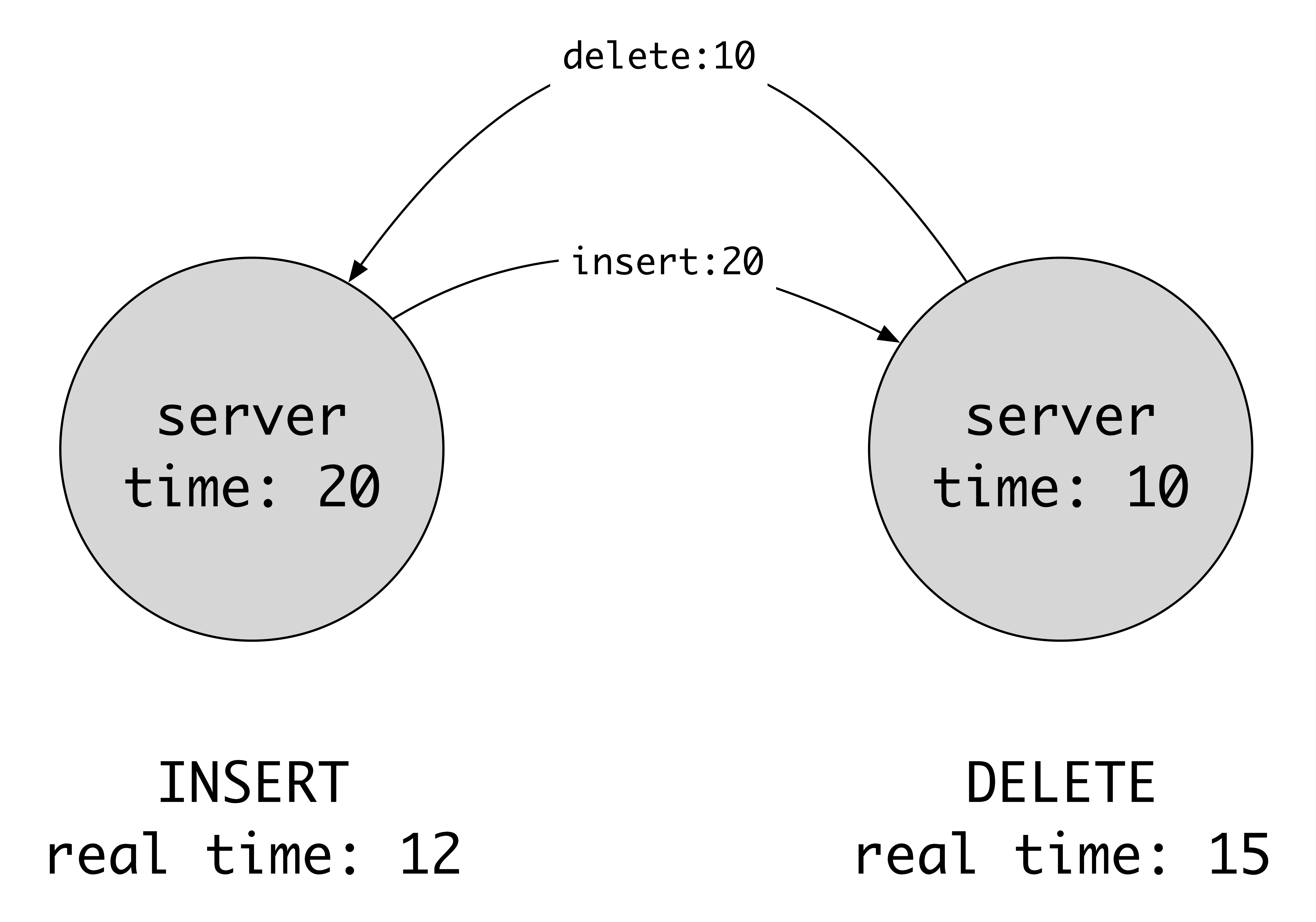
0 to **20** of 100 available for paging

@timestamp	@severity	@message
2013-06-27T09:56:24.689917+03:00	notice	this is a test2
2013-06-27T09:56:20.685410+03:00	notice	this is a test4
2013-06-27T09:56:20.681484+03:00	notice	this is a test0
2013-06-27T09:56:20.674361+03:00	notice	this is a test0

Gotchas

Incorrect Server Times

- Everything is written with a timestamp
- Last write wins
- Usually supplied by coordinator
- Can also be supplied by client
- What if your timestamps are wrong because your clocks are off?
- Always install ntpd!



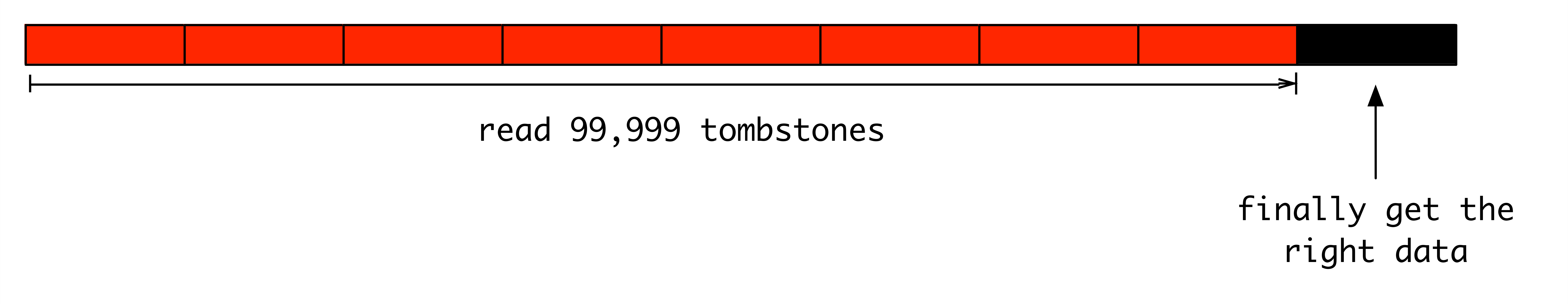
Tombstones

- Tombstones are a marker that data no longer exists
- Tombstones have a timestamp just like normal data
- They say "at time X, this no longer exists"

NO DATA HERE
PK=1
1411069925647419
www.tombstonebuilder.com

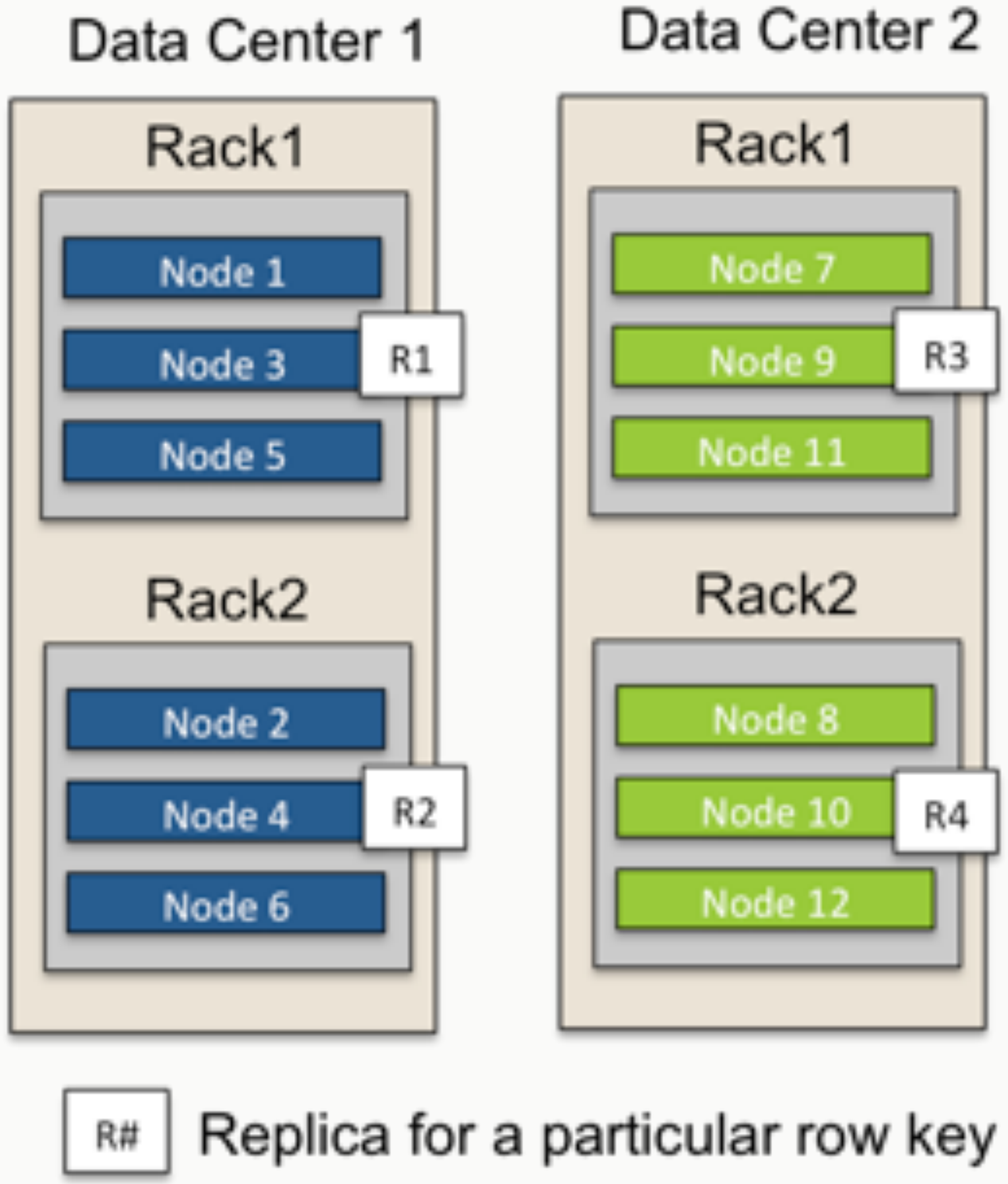
Tombstone Hell

- Queries on partitions with a lot of tombstones require a lot of filtering
- This can be reaaaaaaally slow
- Consider:
  - 100,000 rows in a partition
  - 99,999 are tombstones
  - How long to get a single row?
- Cassandra is not a queue!



Not using a Snitch

- Snitch lets us distribute data in a fault tolerant way
- Changing this with a large cluster is time consuming
- Dynamic Snitching
  - use the fastest replica for reads
- RackInferring (uses IP to pick replicas)
- DC aware
- PropertyFileSnitch (cassandra-topology.properties)
- EC2Snitch & EC2MultiRegion
- GoogleCloudSnitch
- GossipingPropertyFileSnitch (recommended)



Version Mismatch

- SSTable format changed between versions, making streaming incompatible
- Version mismatch can break bootstrap, repair, and decommission
- Introducing new nodes? Stick w/ the same version
- Upgrade nodes in place
  - One at a time
  - One rack / AZ at a time (requires proper snitch)

Disk Space not Reclaimed

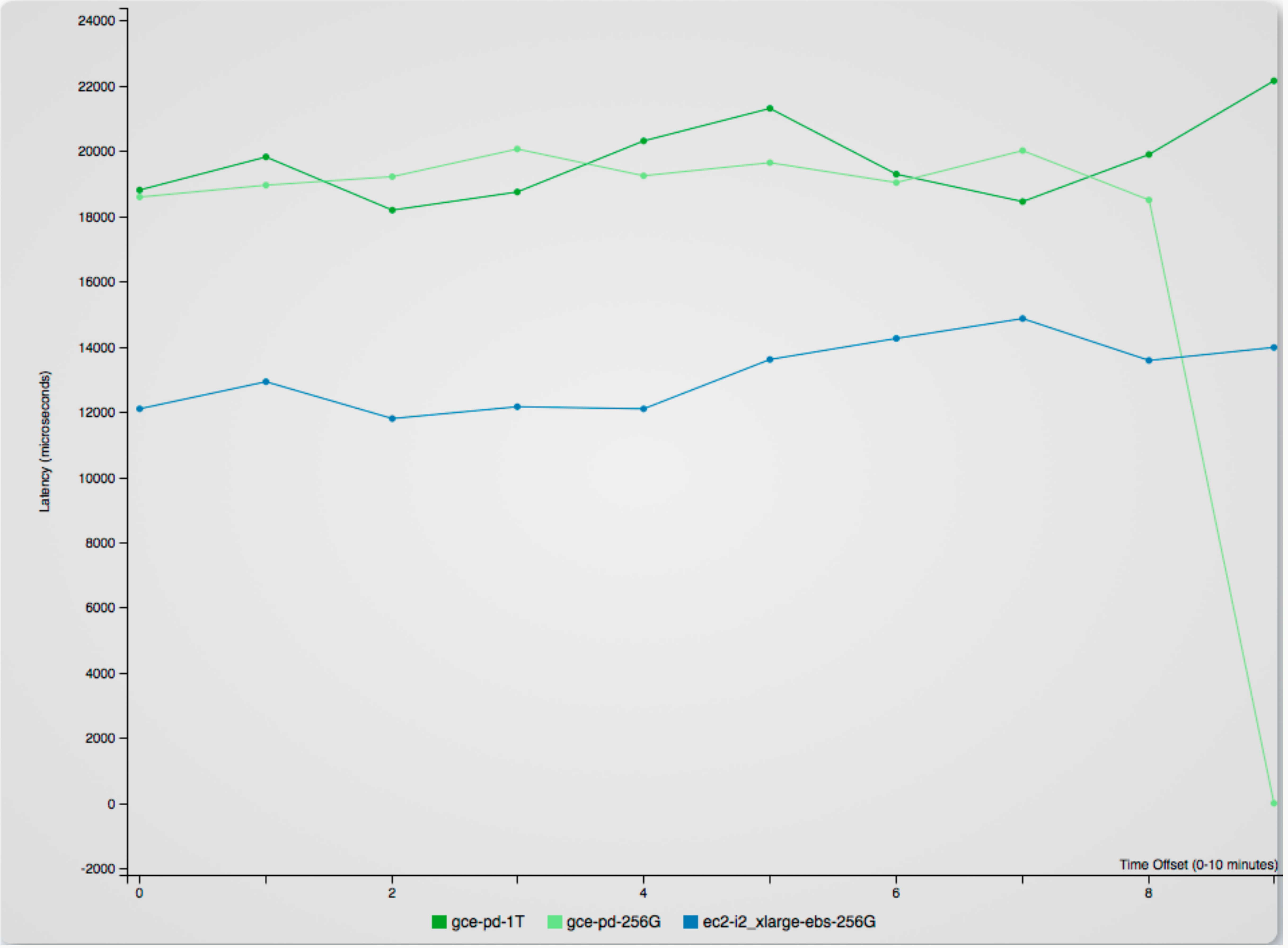
- When you add new nodes, data is streamed from existing nodes
- … but it's not deleted from them after
- You need to run a nodetool cleanup
- Otherwise you'll run out of space just by adding nodes

Low Disk Space

You are running out of disk space on Local Disk (C:). To free space on this drive by deleting old or unnecessary files, click here...

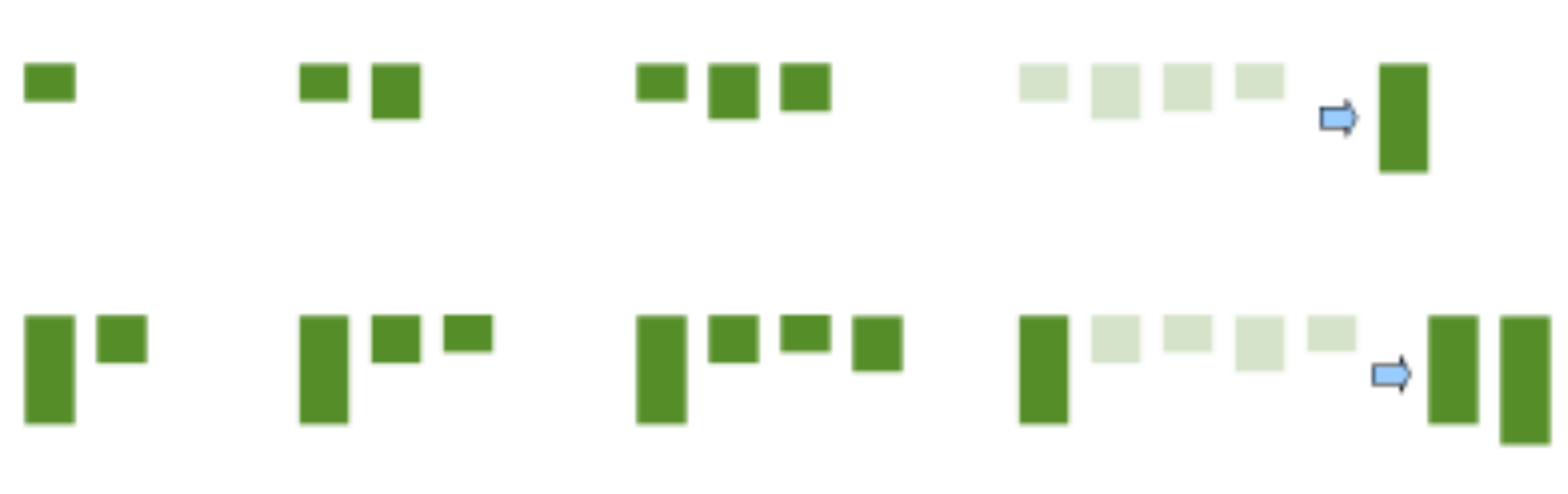
Using Shared Storage

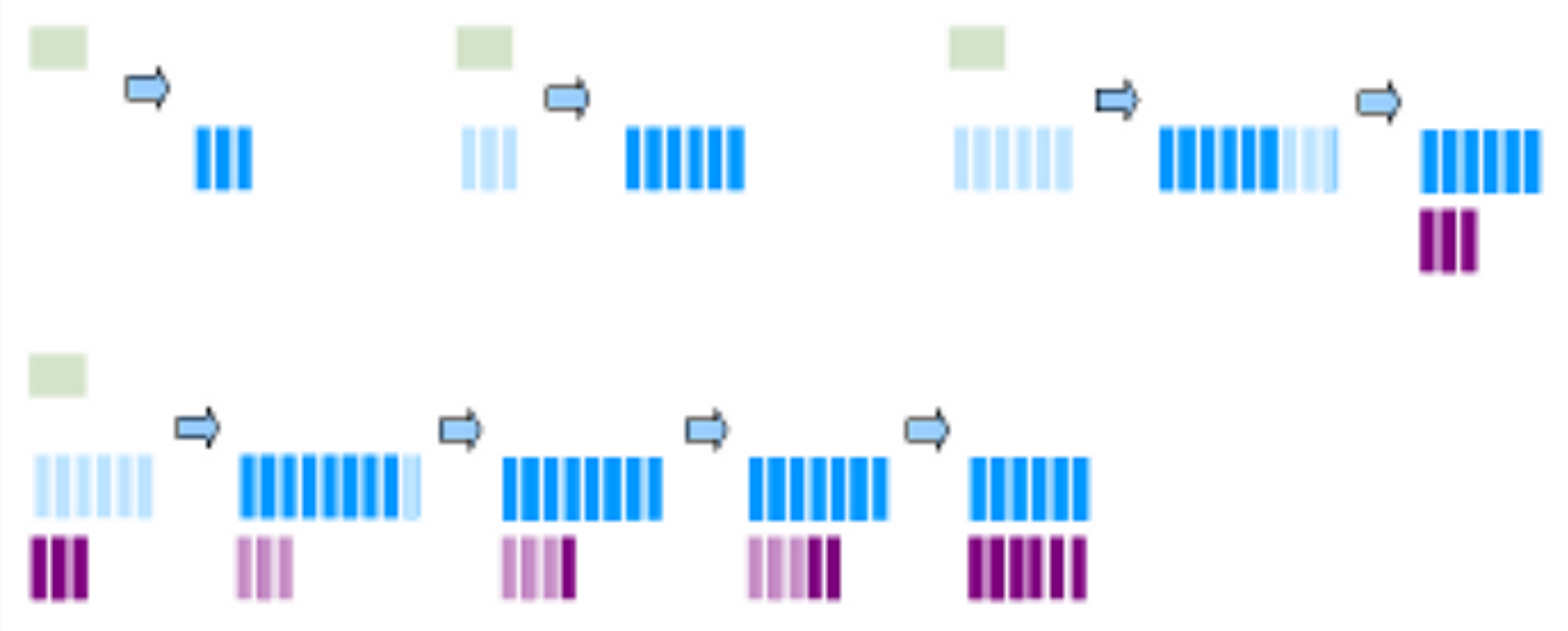
- Single point of failure
- High latency
- Expensive
- Performance is about latency
- Can increase throughput with more disks
- In general avoid EBS, SAN, NAS



Compaction

- Compaction merges SSTables
- Too much compaction?
- Opscenter provides insight into compaction cluster wide
- nodetool
  - compactionhistory
  - getcompactionthroughput
- Leveled vs Size Tiered vs Date Tiered
  - Leveled on SSD + Read Heavy
  - Size tiered on Spinning rust
  - Size tiered is great for write heavy time series workloads
  - Date tiered is new and is showing HUGE promise

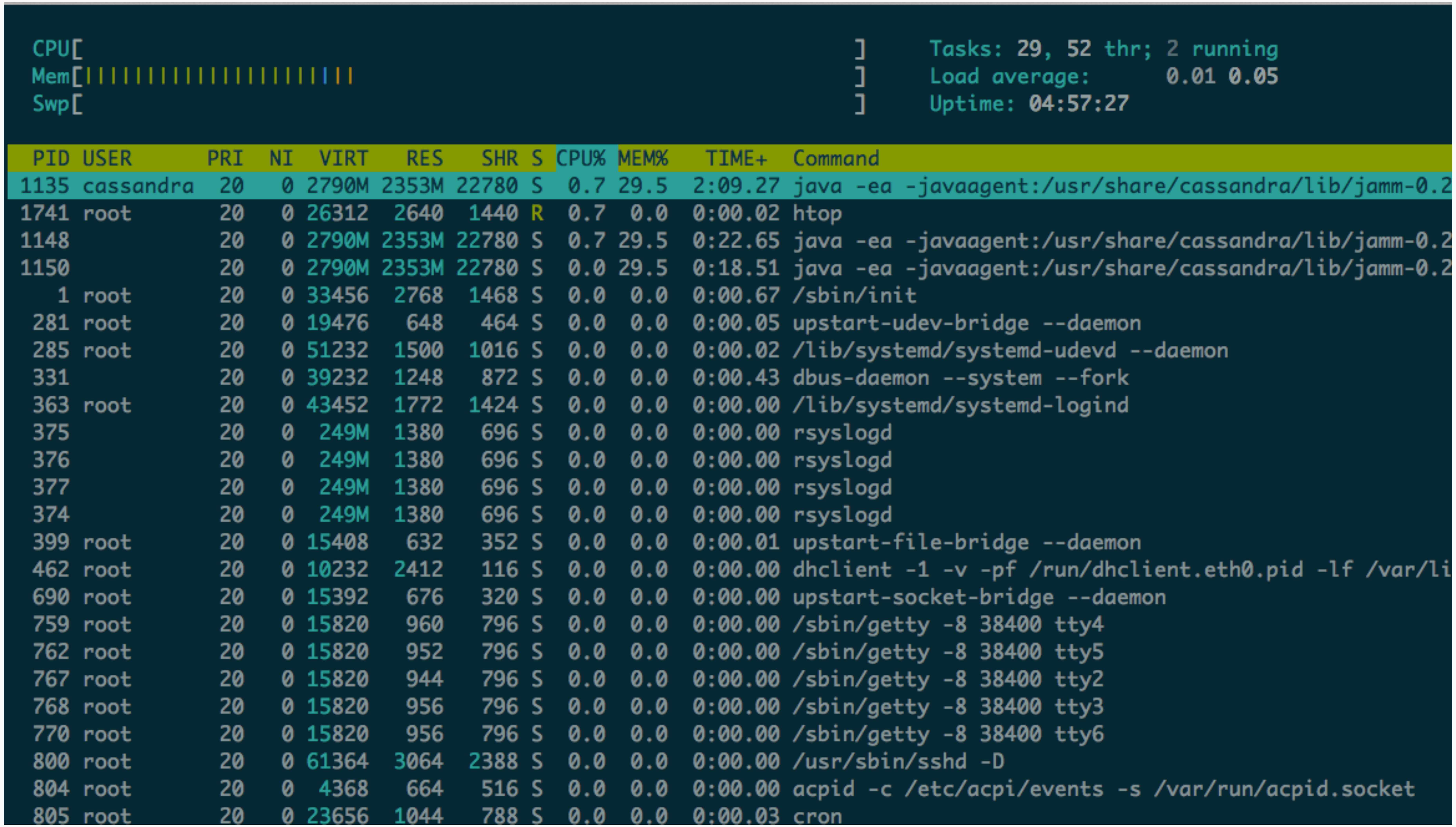




Diagnostic Tools

htop

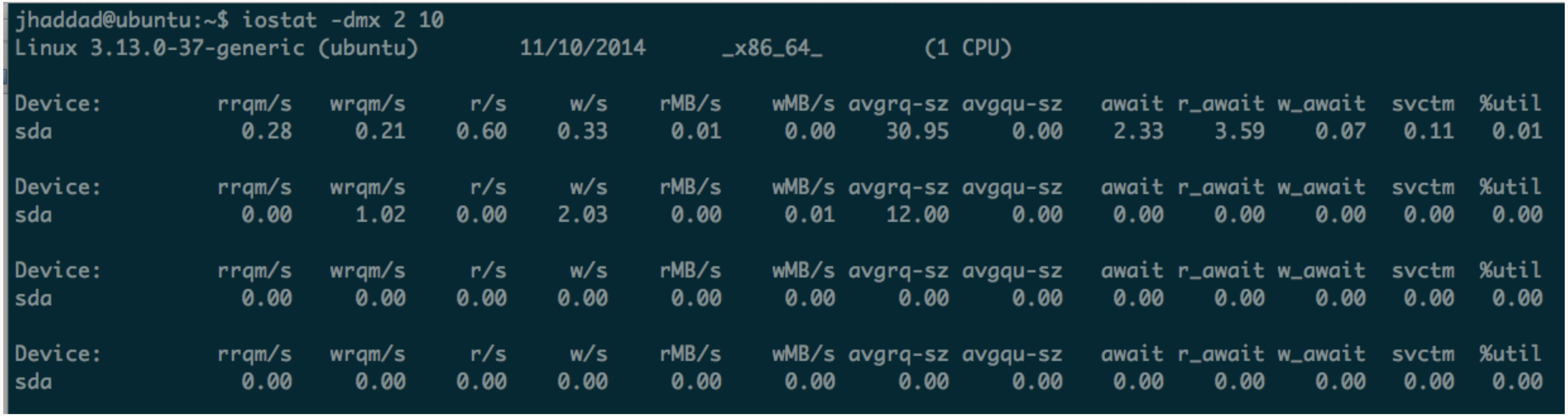
- Process overview - nicer than top



```
 CPU[                                                              ]     Tasks: 29, 52 thr; 2 running
 Mem[||||||||||||||||||||||                                        ]     Load average:      0.01 0.05
 Swp[                                                              ]     Uptime: 04:57:27

 PID USER      PRI  NI  VIRT   RES   SHR S CPU% MEM%   TIME+  Command
1135 cassandra  20   0 2790M 2353M 22780 S  0.7 29.5  2:09.27 java -ea -javaagent:/usr/share/cassandra/lib/jamm-0.2
1741 root       20   0 26312  2640  1440 R  0.7  0.0  0:00.02 htop
1148            20   0 2790M 2353M 22780 S  0.7 29.5  0:22.65 java -ea -javaagent:/usr/share/cassandra/lib/jamm-0.2
1150            20   0 2790M 2353M 22780 S  0.0 29.5  0:18.51 java -ea -javaagent:/usr/share/cassandra/lib/jamm-0.2
   1 root       20   0 33456  2768  1468 S  0.0  0.0  0:00.67 /sbin/init
 281 root       20   0 19476   648   464 S  0.0  0.0  0:00.05 upstart-udev-bridge --daemon
 285 root       20   0 51232  1500  1016 S  0.0  0.0  0:00.02 /lib/systemd/systemd-udevd --daemon
 331            20   0 39232  1248   872 S  0.0  0.0  0:00.43 dbus-daemon --system --fork
 363 root       20   0 43452  1772  1424 S  0.0  0.0  0:00.00 /lib/systemd/systemd-logind
 375            20   0  249M  1380   696 S  0.0  0.0  0:00.00 rsyslogd
 376            20   0  249M  1380   696 S  0.0  0.0  0:00.00 rsyslogd
 377            20   0  249M  1380   696 S  0.0  0.0  0:00.00 rsyslogd
 374            20   0  249M  1380   696 S  0.0  0.0  0:00.00 rsyslogd
 399 root       20   0 15408   632   352 S  0.0  0.0  0:00.01 upstart-file-bridge --daemon
 462 root       20   0 10232  2412   116 S  0.0  0.0  0:00.00 dhclient -1 -v -pf /run/dhclient.eth0.pid -lf /var/li
 690 root       20   0 15392   676   320 S  0.0  0.0  0:00.00 upstart-socket-bridge --daemon
 759 root       20   0 15820   960   796 S  0.0  0.0  0:00.00 /sbin/getty -8 38400 tty4
 762 root       20   0 15820   952   796 S  0.0  0.0  0:00.00 /sbin/getty -8 38400 tty5
 767 root       20   0 15820   944   796 S  0.0  0.0  0:00.00 /sbin/getty -8 38400 tty2
 768 root       20   0 15820   956   796 S  0.0  0.0  0:00.00 /sbin/getty -8 38400 tty3
 770 root       20   0 15820   956   796 S  0.0  0.0  0:00.00 /sbin/getty -8 38400 tty6
 800 root       20   0 61364  3064  2388 S  0.0  0.0  0:00.00 /usr/sbin/sshd -D
 804 root       20   0  4368   664   516 S  0.0  0.0  0:00.00 acpid -c /etc/acpi/events -s /var/run/acpid.socket
 805 root       20   0 23656  1044   788 S  0.0  0.0  0:00.03 cron
```

iostat

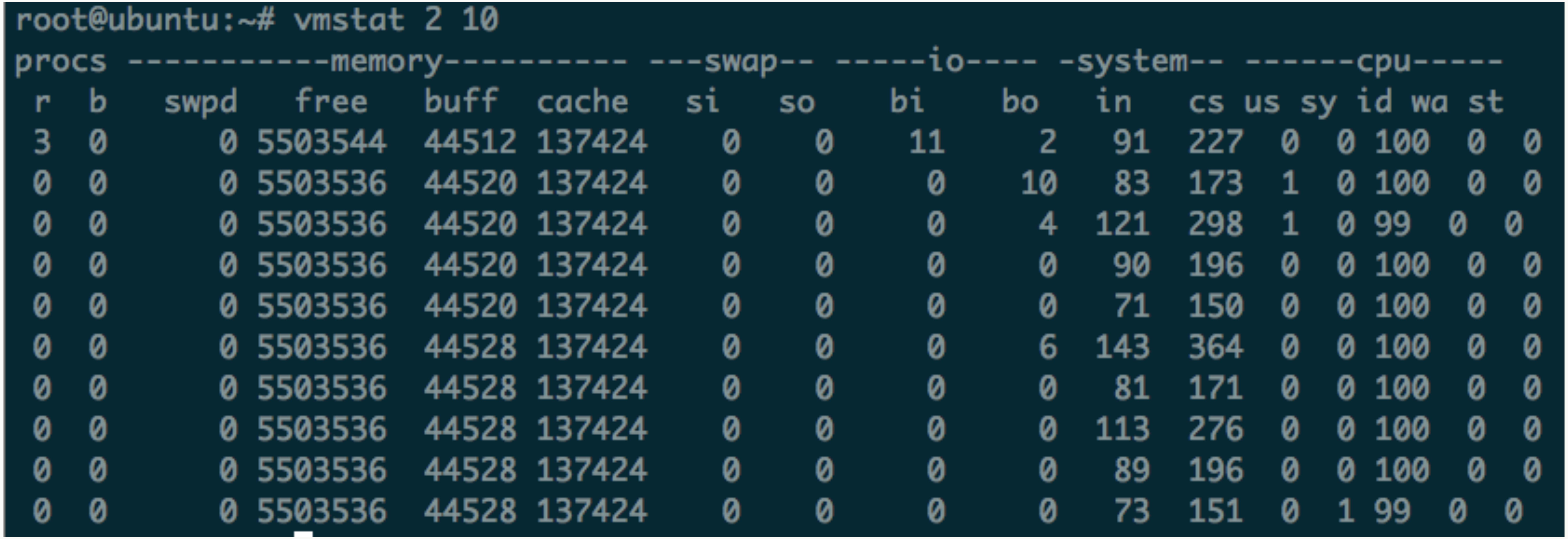
- Disk stats
  - Queue size, wait times
- Ignore %util



```
jhaddad@ubuntu:~$ iostat -dmx 2 10
Linux 3.13.0-37-generic (ubuntu)       11/10/2014      _x86_64_        (1 CPU)

Device:         rrqm/s   wrqm/s     r/s     w/s    rMB/s    wMB/s avgrq-sz avgqu-sz   await r_await w_await  svctm  %util
sda               0.28     0.21    0.60    0.33     0.01     0.00    30.95     0.00    2.33    3.59    0.07   0.11   0.01

Device:         rrqm/s   wrqm/s     r/s     w/s    rMB/s    wMB/s avgrq-sz avgqu-sz   await r_await w_await  svctm  %util
sda               0.00     1.02    0.00    2.03     0.00     0.01    12.00     0.00    0.00    0.00    0.00   0.00   0.00

Device:         rrqm/s   wrqm/s     r/s     w/s    rMB/s    wMB/s avgrq-sz avgqu-sz   await r_await w_await  svctm  %util
sda               0.00     0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00    0.00    0.00   0.00   0.00

Device:         rrqm/s   wrqm/s     r/s     w/s    rMB/s    wMB/s avgrq-sz avgqu-sz   await r_await w_await  svctm  %util
sda               0.00     0.00    0.00    0.00     0.00     0.00     0.00     0.00    0.00    0.00    0.00   0.00   0.00
```

vmstat

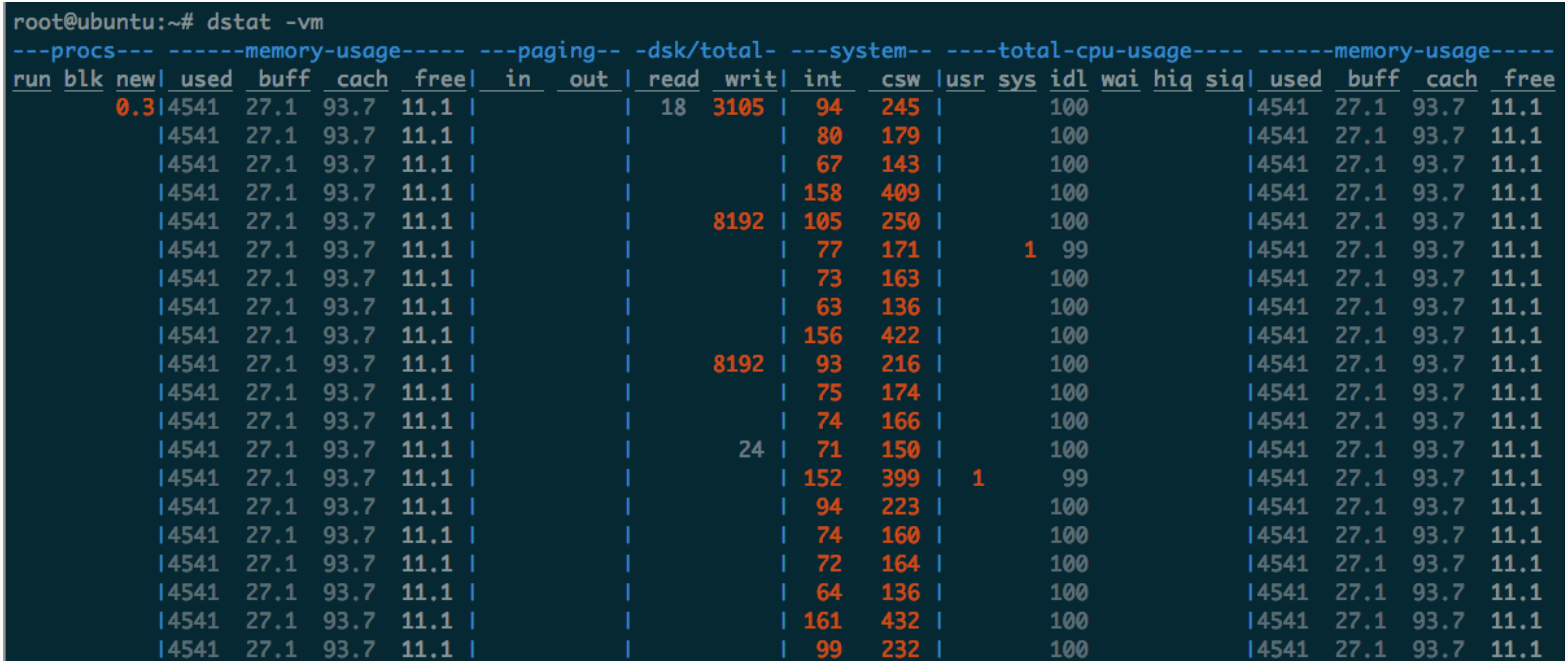
- virtual memory statistics
- Am I swapping?
- Reports at an interval, to an optional count



```
root@ubuntu:~# vmstat 2 10
procs -----------memory---------- ---swap-- -----io---- -system-- ------cpu-----
 r  b   swpd   free   buff  cache   si   so    bi    bo   in   cs us sy id wa st
 3  0      0 5503544  44512 137424    0    0    11     2   91  227  0  0 100  0  0
 0  0      0 5503536  44520 137424    0    0     0    10   83  173  1  0 100  0  0
 0  0      0 5503536  44520 137424    0    0     0     4  121  298  1  0 99  0  0
 0  0      0 5503536  44520 137424    0    0     0     0   90  196  0  0 100  0  0
 0  0      0 5503536  44520 137424    0    0     0     0   71  150  0  0 100  0  0
 0  0      0 5503536  44528 137424    0    0     0     6  143  364  0  0 100  0  0
 0  0      0 5503536  44528 137424    0    0     0     0   81  171  0  0 100  0  0
 0  0      0 5503536  44528 137424    0    0     0     0  113  276  0  0 100  0  0
 0  0      0 5503536  44528 137424    0    0     0     0   89  196  0  0 100  0  0
 0  0      0 5503536  44528 137424    0    0     0     0   73  151  0  1 99  0  0
```

dstat

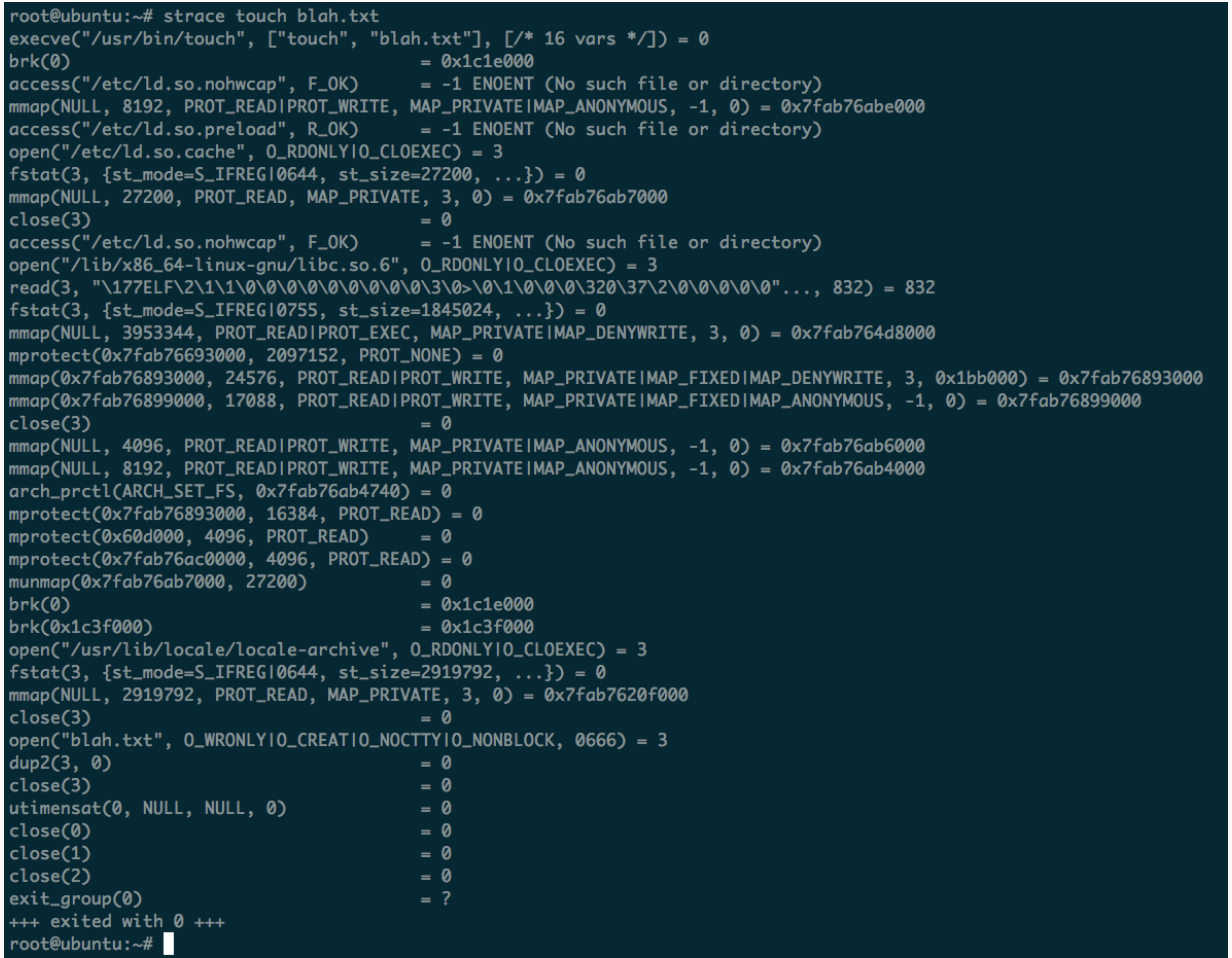
- Flexible look at network, CPU, memory, disk



```
root@ubuntu:~# dstat -vm
---procs--- ------memory-usage----- ---paging-- -dsk/total- ---system-- ----total-cpu-usage---- ------memory-usage-----
run blk new| used  buff  cach  free|  in   out | read  writ| int   csw |usr sys idl wai hiq siq| used  buff  cach  free
        0.3|4541  27.1  93.7  11.1 |           |  18  3105 |  94   245 |        100            |4541  27.1  93.7  11.1
           |4541  27.1  93.7  11.1 |           |           |  80   179 |        100            |4541  27.1  93.7  11.1
           |4541  27.1  93.7  11.1 |           |           |  67   143 |        100            |4541  27.1  93.7  11.1
           |4541  27.1  93.7  11.1 |           |           | 158   409 |        100            |4541  27.1  93.7  11.1
           |4541  27.1  93.7  11.1 |           |      8192 | 105   250 |        100            |4541  27.1  93.7  11.1
           |4541  27.1  93.7  11.1 |           |           |  77   171 |      1  99            |4541  27.1  93.7  11.1
           |4541  27.1  93.7  11.1 |           |           |  73   163 |        100            |4541  27.1  93.7  11.1
           |4541  27.1  93.7  11.1 |           |           |  63   136 |        100            |4541  27.1  93.7  11.1
           |4541  27.1  93.7  11.1 |           |           | 156   422 |        100            |4541  27.1  93.7  11.1
           |4541  27.1  93.7  11.1 |           |      8192 |  93   216 |        100            |4541  27.1  93.7  11.1
           |4541  27.1  93.7  11.1 |           |           |  75   174 |        100            |4541  27.1  93.7  11.1
           |4541  27.1  93.7  11.1 |           |           |  74   166 |        100            |4541  27.1  93.7  11.1
           |4541  27.1  93.7  11.1 |           |        24 |  71   150 |        100            |4541  27.1  93.7  11.1
           |4541  27.1  93.7  11.1 |           |           | 152   399 |  1      99            |4541  27.1  93.7  11.1
           |4541  27.1  93.7  11.1 |           |           |  94   223 |        100            |4541  27.1  93.7  11.1
           |4541  27.1  93.7  11.1 |           |           |  74   160 |        100            |4541  27.1  93.7  11.1
           |4541  27.1  93.7  11.1 |           |           |  72   164 |        100            |4541  27.1  93.7  11.1
           |4541  27.1  93.7  11.1 |           |           |  64   136 |        100            |4541  27.1  93.7  11.1
           |4541  27.1  93.7  11.1 |           |           | 161   432 |        100            |4541  27.1  93.7  11.1
           |4541  27.1  93.7  11.1 |           |           |  99   232 |        100            |4541  27.1  93.7  11.1
```

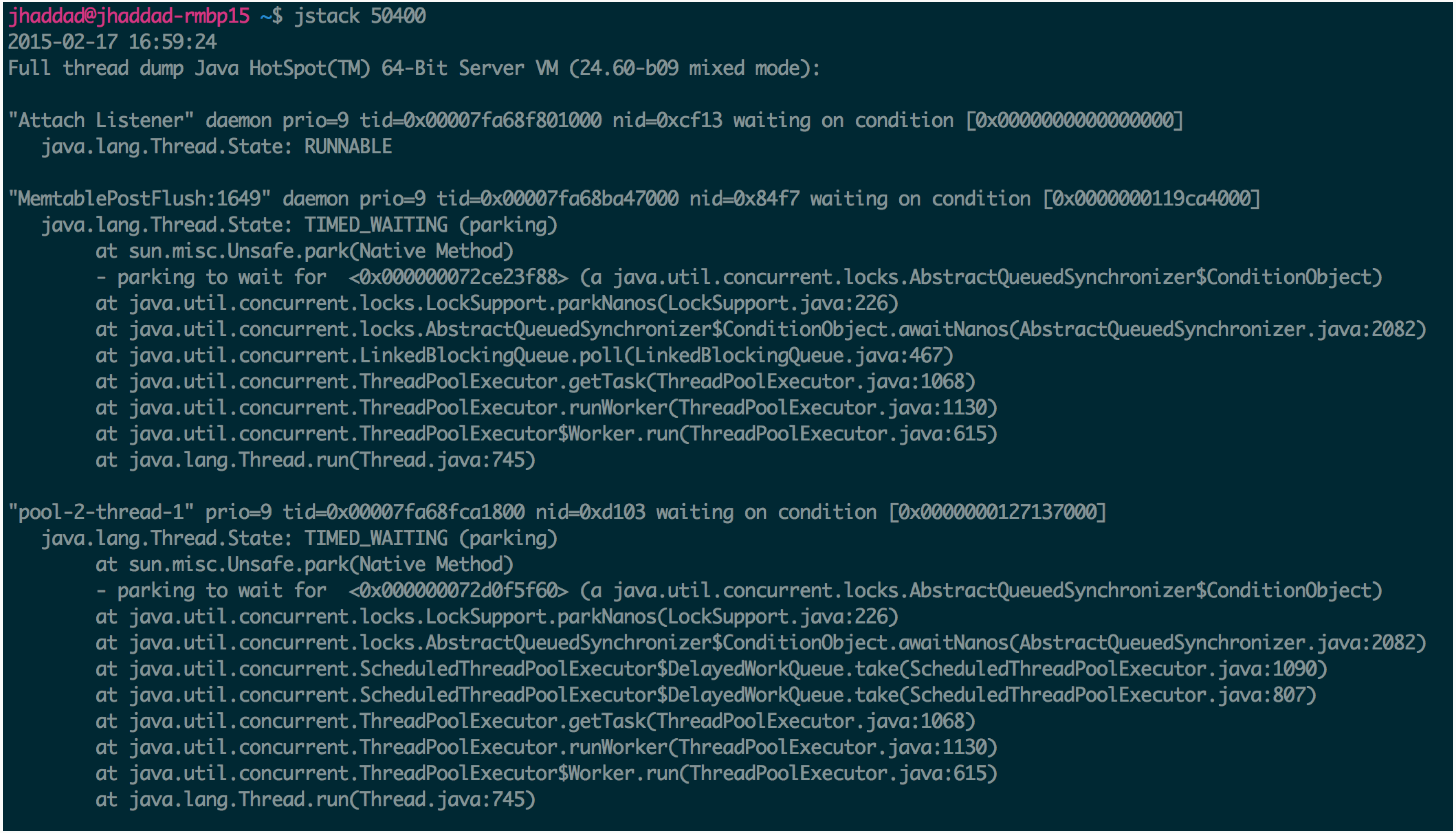
strace

- What is my process doing?
- See all system calls
- Filterable with -e
- Can attach to running processes



```
root@ubuntu:~# strace touch blah.txt
execve("/usr/bin/touch", ["touch", "blah.txt"], [/* 16 vars */]) = 0
brk(0)                                  = 0x1c1e000
access("/etc/ld.so.nohwcap", F_OK)      = -1 ENOENT (No such file or directory)
mmap(NULL, 8192, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) = 0x7fab76abe000
access("/etc/ld.so.preload", R_OK)      = -1 ENOENT (No such file or directory)
open("/etc/ld.so.cache", O_RDONLY|O_CLOEXEC) = 3
fstat(3, {st_mode=S_IFREG|0644, st_size=27200, ...}) = 0
mmap(NULL, 27200, PROT_READ, MAP_PRIVATE, 3, 0) = 0x7fab76ab7000
close(3)                                = 0
access("/etc/ld.so.nohwcap", F_OK)      = -1 ENOENT (No such file or directory)
open("/lib/x86_64-linux-gnu/libc.so.6", O_RDONLY|O_CLOEXEC) = 3
read(3, "\177ELF\2\1\1\0\0\0\0\0\0\0\0\0\3\0>\0\1\0\0\0\320\37\2\0\0\0\0\0"..., 832) = 832
fstat(3, {st_mode=S_IFREG|0755, st_size=1845024, ...}) = 0
mmap(NULL, 3953344, PROT_READ|PROT_EXEC, MAP_PRIVATE|MAP_DENYWRITE, 3, 0) = 0x7fab764d8000
mprotect(0x7fab76693000, 2097152, PROT_NONE) = 0
mmap(0x7fab76893000, 24576, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_FIXED|MAP_DENYWRITE, 3, 0x1bb000) = 0x7fab76893000
mmap(0x7fab76899000, 17088, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_FIXED|MAP_ANONYMOUS, -1, 0) = 0x7fab76899000
close(3)                                = 0
mmap(NULL, 4096, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) = 0x7fab76ab6000
mmap(NULL, 8192, PROT_READ|PROT_WRITE, MAP_PRIVATE|MAP_ANONYMOUS, -1, 0) = 0x7fab76ab4000
arch_prctl(ARCH_SET_FS, 0x7fab76ab4740) = 0
mprotect(0x7fab76893000, 16384, PROT_READ) = 0
mprotect(0x60d000, 4096, PROT_READ)     = 0
mprotect(0x7fab76ac0000, 4096, PROT_READ) = 0
munmap(0x7fab76ab7000, 27200)           = 0
brk(0)                                  = 0x1c1e000
brk(0x1c3f000)                          = 0x1c3f000
open("/usr/lib/locale/locale-archive", O_RDONLY|O_CLOEXEC) = 3
fstat(3, {st_mode=S_IFREG|0644, st_size=2919792, ...}) = 0
mmap(NULL, 2919792, PROT_READ, MAP_PRIVATE, 3, 0) = 0x7fab7620f000
close(3)                                = 0
open("blah.txt", O_WRONLY|O_CREAT|O_NOCTTY|O_NONBLOCK, 0666) = 3
dup2(3, 0)                              = 0
close(3)                                = 0
utimensat(0, NULL, NULL, 0)             = 0
close(0)                                = 0
close(1)                                = 0
close(2)                                = 0
exit_group(0)                           = ?
+++ exited with 0 +++
root@ubuntu:~# 
```

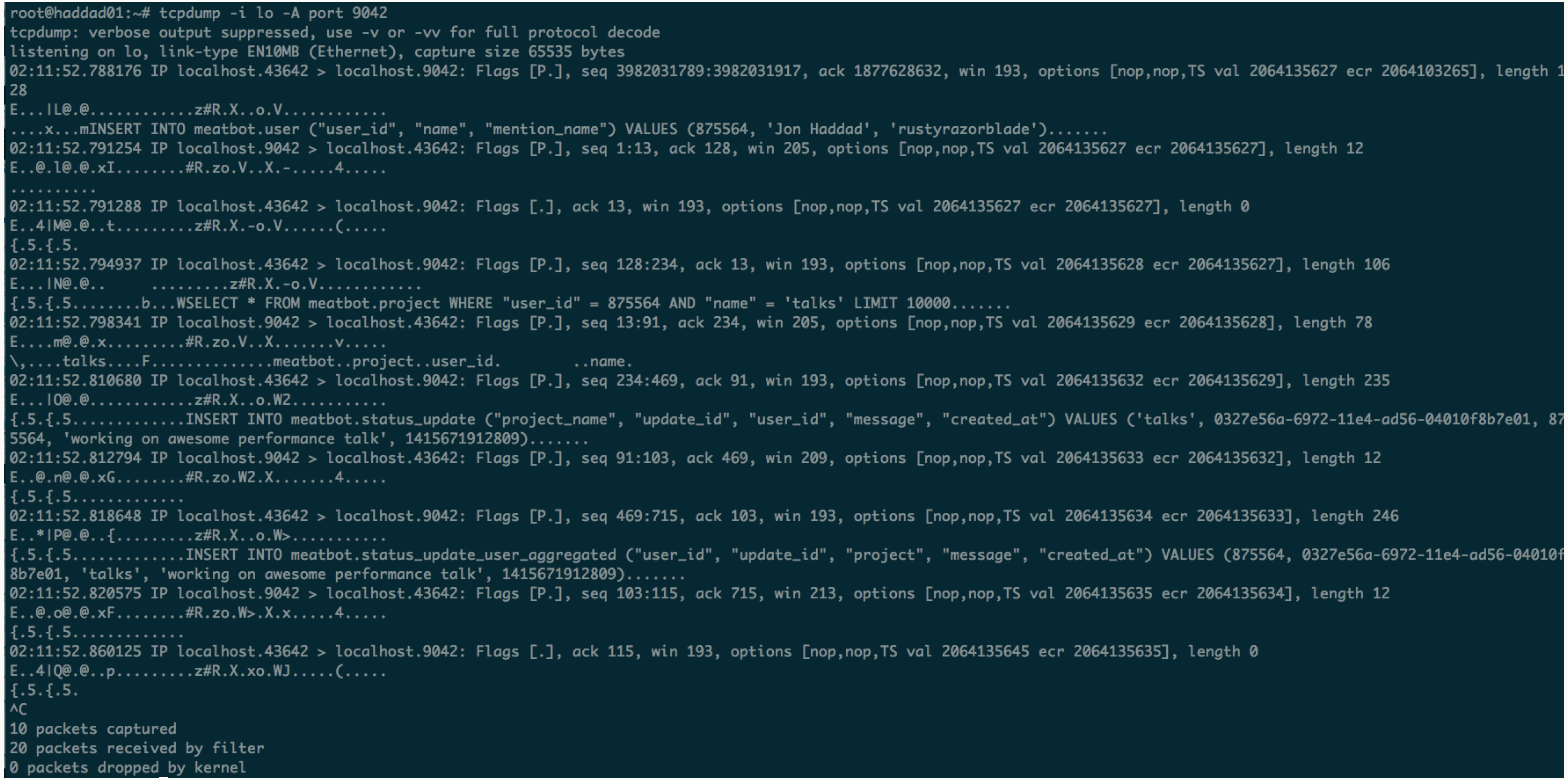
jstack



```
jhaddad@jhaddad-rmbp15 ~$ jstack 50400
2015-02-17 16:59:24
Full thread dump Java HotSpot(TM) 64-Bit Server VM (24.60-b09 mixed mode):

"Attach Listener" daemon prio=9 tid=0x00007fa68f801000 nid=0xcf13 waiting on condition [0x0000000000000000]
   java.lang.Thread.State: RUNNABLE

"MemtablePostFlush:1649" daemon prio=9 tid=0x00007fa68ba47000 nid=0x84f7 waiting on condition [0x0000000119ca4000]
   java.lang.Thread.State: TIMED_WAITING (parking)
        at sun.misc.Unsafe.park(Native Method)
        - parking to wait for  <0x000000072ce23f88> (a java.util.concurrent.locks.AbstractQueuedSynchronizer$ConditionObject)
        at java.util.concurrent.locks.LockSupport.parkNanos(LockSupport.java:226)
        at java.util.concurrent.locks.AbstractQueuedSynchronizer$ConditionObject.awaitNanos(AbstractQueuedSynchronizer.java:2082)
        at java.util.concurrent.LinkedBlockingQueue.poll(LinkedBlockingQueue.java:467)
        at java.util.concurrent.ThreadPoolExecutor.getTask(ThreadPoolExecutor.java:1068)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1130)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:615)
        at java.lang.Thread.run(Thread.java:745)

"pool-2-thread-1" prio=9 tid=0x00007fa68fca1800 nid=0xd103 waiting on condition [0x0000000127137000]
   java.lang.Thread.State: TIMED_WAITING (parking)
        at sun.misc.Unsafe.park(Native Method)
        - parking to wait for  <0x000000072d0f5f60> (a java.util.concurrent.locks.AbstractQueuedSynchronizer$ConditionObject)
        at java.util.concurrent.locks.LockSupport.parkNanos(LockSupport.java:226)
        at java.util.concurrent.locks.AbstractQueuedSynchronizer$ConditionObject.awaitNanos(AbstractQueuedSynchronizer.java:2082)
        at java.util.concurrent.ScheduledThreadPoolExecutor$DelayedWorkQueue.take(ScheduledThreadPoolExecutor.java:1090)
        at java.util.concurrent.ScheduledThreadPoolExecutor$DelayedWorkQueue.take(ScheduledThreadPoolExecutor.java:807)
        at java.util.concurrent.ThreadPoolExecutor.getTask(ThreadPoolExecutor.java:1068)
        at java.util.concurrent.ThreadPoolExecutor.runWorker(ThreadPoolExecutor.java:1130)
        at java.util.concurrent.ThreadPoolExecutor$Worker.run(ThreadPoolExecutor.java:615)
        at java.lang.Thread.run(Thread.java:745)
```

tcpdump

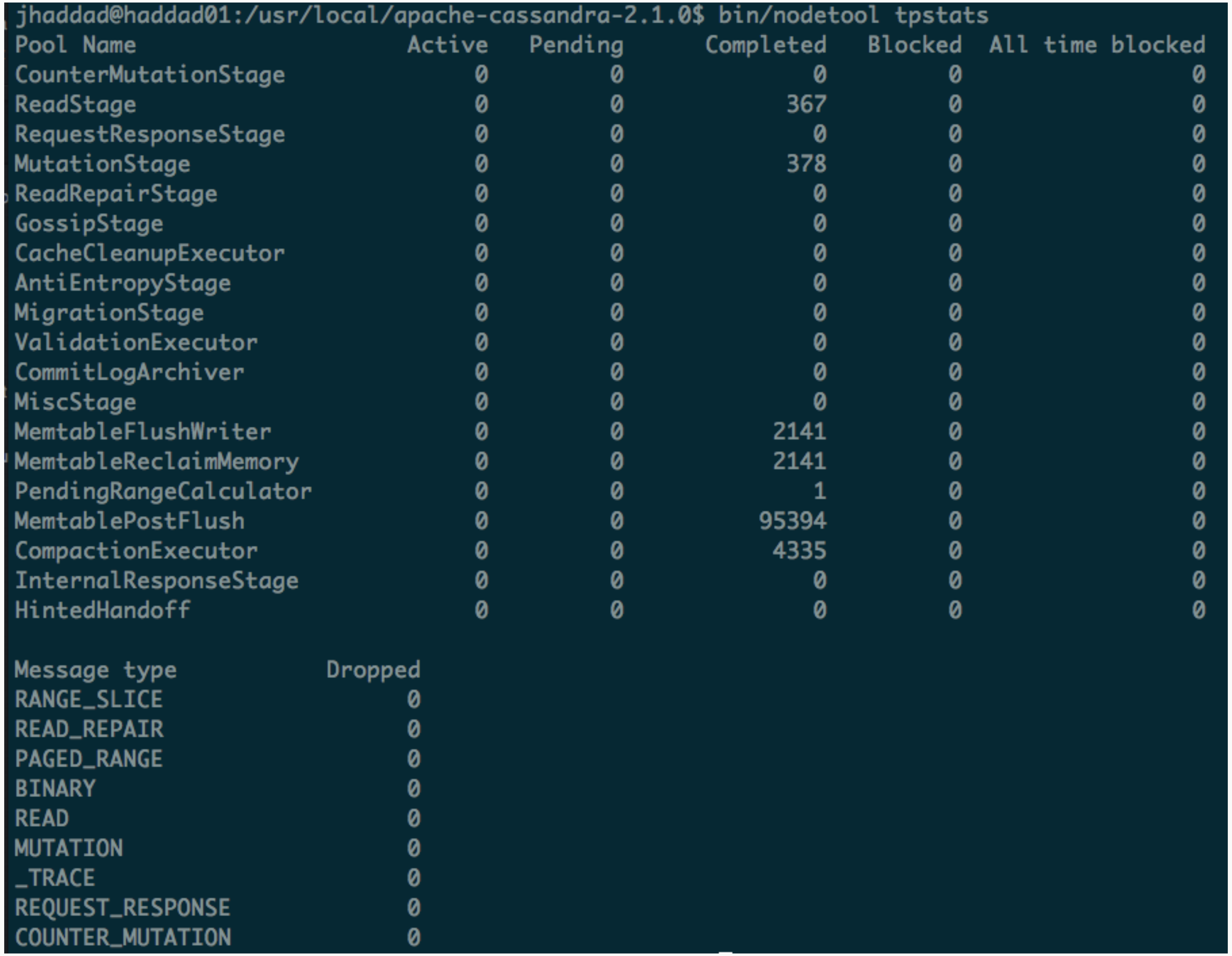
- Watch network traffic



```
root@haddad01:~# tcpdump -i lo -A port 9042
tcpdump: verbose output suppressed, use -v or -vv for full protocol decode
listening on lo, link-type EN10MB (Ethernet), capture size 65535 bytes
02:11:52.788176 IP localhost.43642 > localhost.9042: Flags [P.], seq 3982031789:3982031917, ack 1877628632, win 193, options [nop,nop,TS val 2064135627 ecr 2064103265], length 1
28
E...|L@.@............z#R.X..o.V............
....x...mINSERT INTO meatbot.user ("user_id", "name", "mention_name") VALUES (875564, 'Jon Haddad', 'rustyrazorblade').......
02:11:52.791254 IP localhost.9042 > localhost.43642: Flags [P.], seq 1:13, ack 128, win 205, options [nop,nop,TS val 2064135627 ecr 2064135627], length 12
E..@.l@.@.xI........#R.zo.V..X.-.....4.....
..........
02:11:52.791288 IP localhost.43642 > localhost.9042: Flags [.], ack 13, win 193, options [nop,nop,TS val 2064135627 ecr 2064135627], length 0
E..4|M@.@..t.........z#R.X.-o.V......(.....
{.5.{.5.
02:11:52.794937 IP localhost.43642 > localhost.9042: Flags [P.], seq 128:234, ack 13, win 193, options [nop,nop,TS val 2064135628 ecr 2064135627], length 106
E...|N@.@..     .........z#R.X.-o.V............
{.5.{.5........b...WSELECT * FROM meatbot.project WHERE "user_id" = 875564 AND "name" = 'talks' LIMIT 10000.......
02:11:52.798341 IP localhost.9042 > localhost.43642: Flags [P.], seq 13:91, ack 234, win 205, options [nop,nop,TS val 2064135629 ecr 2064135628], length 78
E....m@.@.x.........#R.zo.V..X.......v.....
\,....talks....F..............meatbot..project..user_id.        ..name.
02:11:52.810680 IP localhost.43642 > localhost.9042: Flags [P.], seq 234:469, ack 91, win 193, options [nop,nop,TS val 2064135632 ecr 2064135629], length 235
E...|O@.@............z#R.X..o.W2..........
{.5.{.5.............INSERT INTO meatbot.status_update ("project_name", "update_id", "user_id", "message", "created_at") VALUES ('talks', 0327e56a-6972-11e4-ad56-04010f8b7e01, 87
5564, 'working on awesome performance talk', 1415671912809).......
02:11:52.812794 IP localhost.9042 > localhost.43642: Flags [P.], seq 91:103, ack 469, win 209, options [nop,nop,TS val 2064135633 ecr 2064135632], length 12
E..@.n@.@.xG........#R.zo.W2.X.......4.....
{.5.{.5.............
02:11:52.818648 IP localhost.43642 > localhost.9042: Flags [P.], seq 469:715, ack 103, win 193, options [nop,nop,TS val 2064135634 ecr 2064135633], length 246
E..*|P@.@..{.........z#R.X..o.W>..........
{.5.{.5.............INSERT INTO meatbot.status_update_user_aggregated ("user_id", "update_id", "project", "message", "created_at") VALUES (875564, 0327e56a-6972-11e4-ad56-04010f
8b7e01, 'talks', 'working on awesome performance talk', 1415671912809).......
02:11:52.820575 IP localhost.9042 > localhost.43642: Flags [P.], seq 103:115, ack 715, win 213, options [nop,nop,TS val 2064135635 ecr 2064135634], length 12
E..@.o@.@.xF........#R.zo.W>.X.x.....4.....
{.5.{.5.............
02:11:52.860125 IP localhost.43642 > localhost.9042: Flags [.], ack 115, win 193, options [nop,nop,TS val 2064135645 ecr 2064135635], length 0
E..4|Q@.@..p.........z#R.X.xo.WJ.....(.....
{.5.{.5.
^C
10 packets captured
20 packets received by filter
0 packets dropped by kernel
```

nodetool tpstats

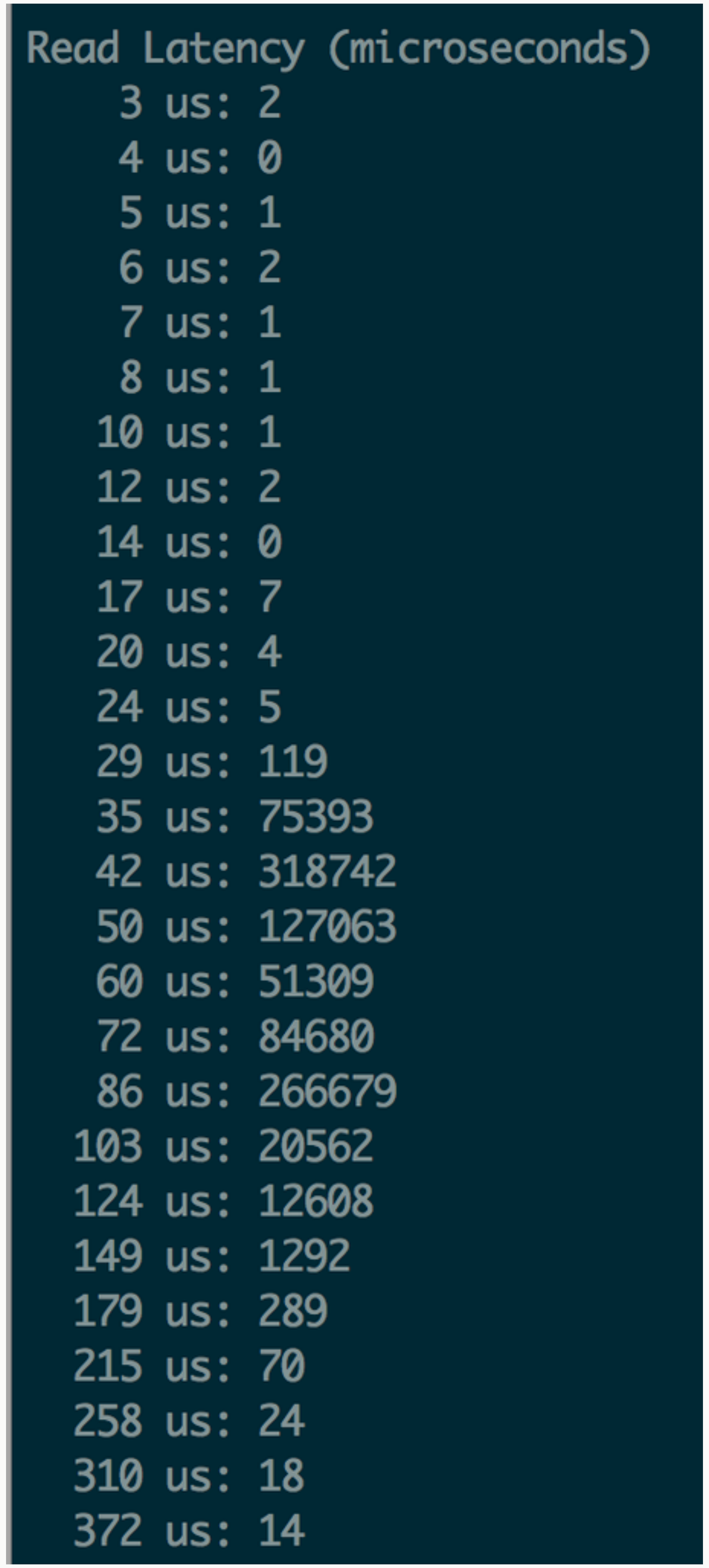
- What's blocked?
- MemtableFlushWriter? - Slow disks!
  - also leads to GC issues
- Dropped mutations?
  - need repair!



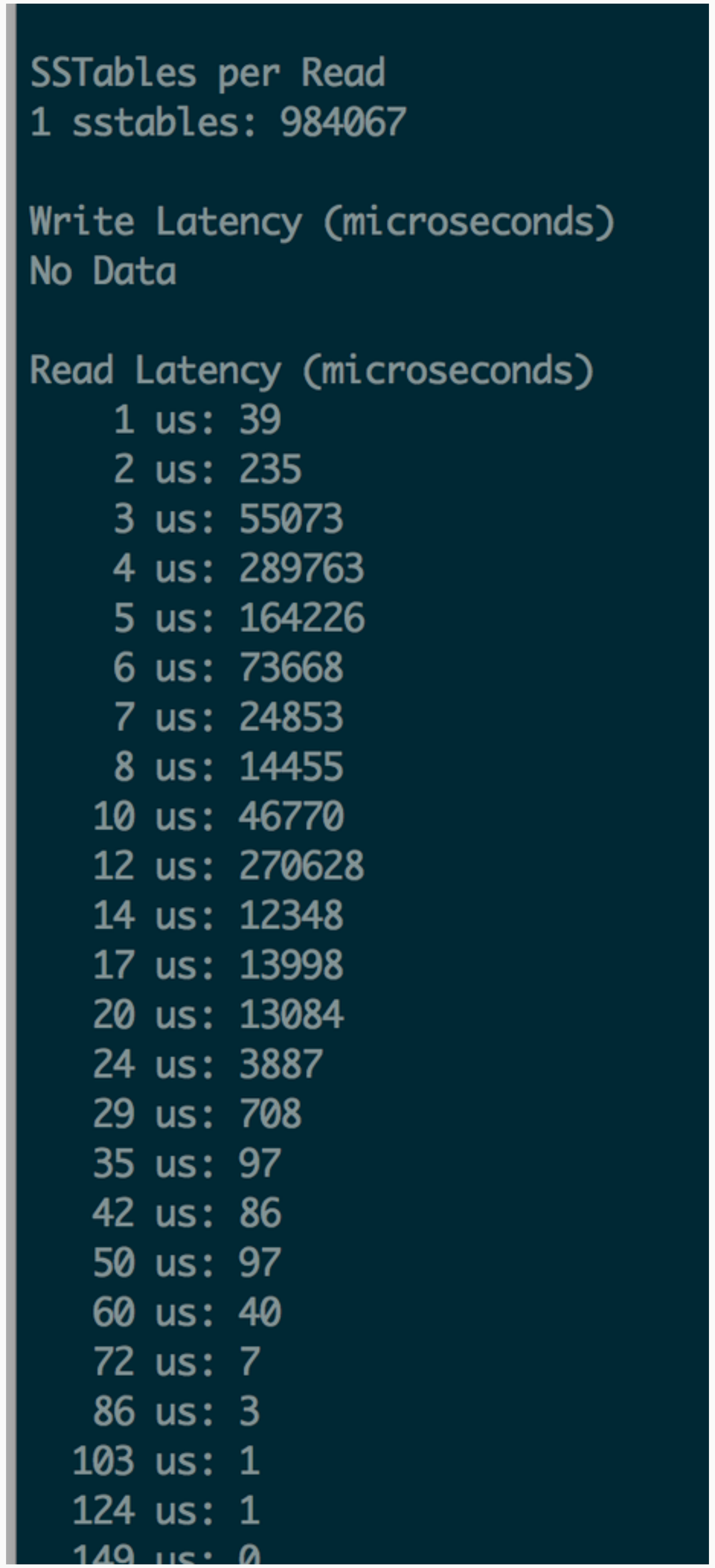
```
jhaddad@haddad01:/usr/local/apache-cassandra-2.1.0$ bin/nodetool tpstats
Pool Name                    Active   Pending      Completed   Blocked  All time blocked
CounterMutationStage              0         0              0         0                 0
ReadStage                         0         0            367         0                 0
RequestResponseStage              0         0              0         0                 0
MutationStage                     0         0            378         0                 0
ReadRepairStage                   0         0              0         0                 0
GossipStage                       0         0              0         0                 0
CacheCleanupExecutor              0         0              0         0                 0
AntiEntropyStage                  0         0              0         0                 0
MigrationStage                    0         0              0         0                 0
ValidationExecutor                0         0              0         0                 0
CommitLogArchiver                 0         0              0         0                 0
MiscStage                         0         0              0         0                 0
MemtableFlushWriter               0         0           2141         0                 0
MemtableReclaimMemory             0         0           2141         0                 0
PendingRangeCalculator            0         0              1         0                 0
MemtablePostFlush                 0         0          95394         0                 0
CompactionExecutor                0         0           4335         0                 0
InternalResponseStage             0         0              0         0                 0
HintedHandoff                     0         0              0         0                 0

Message type           Dropped
RANGE_SLICE                  0
READ_REPAIR                  0
PAGED_RANGE                  0
BINARY                       0
READ                         0
MUTATION                     0
_TRACE                       0
REQUEST_RESPONSE             0
COUNTER_MUTATION             0
```

Histograms

- proxyhistograms
  - High level read and write times
  - Includes network latency
- cfhistograms <keyspace> <table>
  - reports stats for single table on a single node
  - Used to identify tables with performance problems

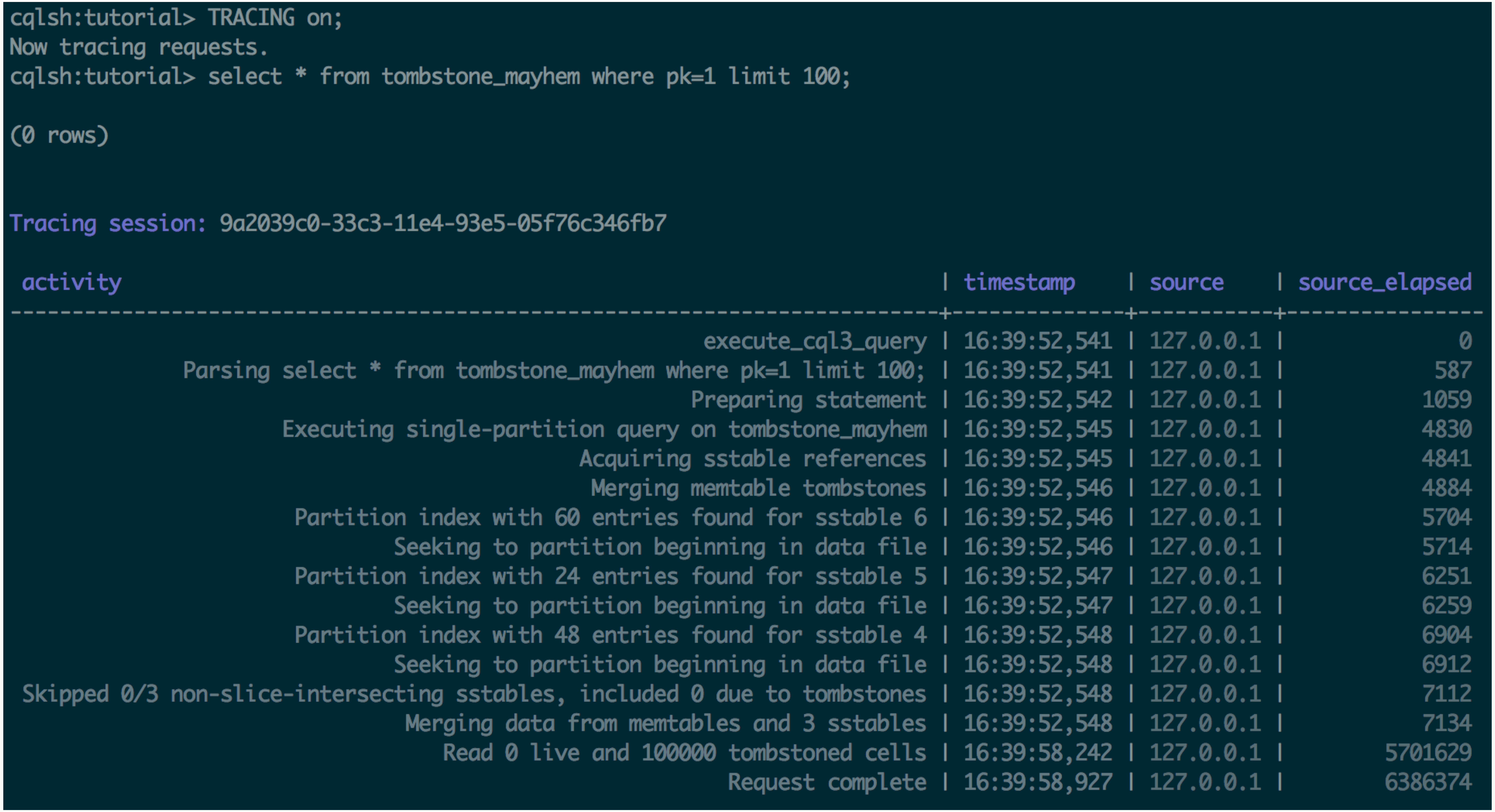


```
Read Latency (microseconds)
      3 us: 2
      4 us: 0
      5 us: 1
      6 us: 2
      7 us: 1
      8 us: 1
     10 us: 1
     12 us: 2
     14 us: 0
     17 us: 7
     20 us: 4
     24 us: 5
     29 us: 119
     35 us: 75393
     42 us: 318742
     50 us: 127063
     60 us: 51309
     72 us: 84680
     86 us: 266679
    103 us: 20562
    124 us: 12608
    149 us: 1292
    179 us: 289
    215 us: 70
    258 us: 24
    310 us: 18
    372 us: 14
```



```
SSTables per Read
1 sstables: 984067

Write Latency (microseconds)
No Data

Read Latency (microseconds)
      1 us: 39
      2 us: 235
      3 us: 55073
      4 us: 289763
      5 us: 164226
      6 us: 73668
      7 us: 24853
      8 us: 14455
     10 us: 46770
     12 us: 270628
     14 us: 12348
     17 us: 13998
     20 us: 13084
     24 us: 3887
     29 us: 708
     35 us: 97
     42 us: 86
     50 us: 97
     60 us: 40
     72 us: 7
     86 us: 3
    103 us: 1
    124 us: 1
    149 us: 0
```

Query Tracing

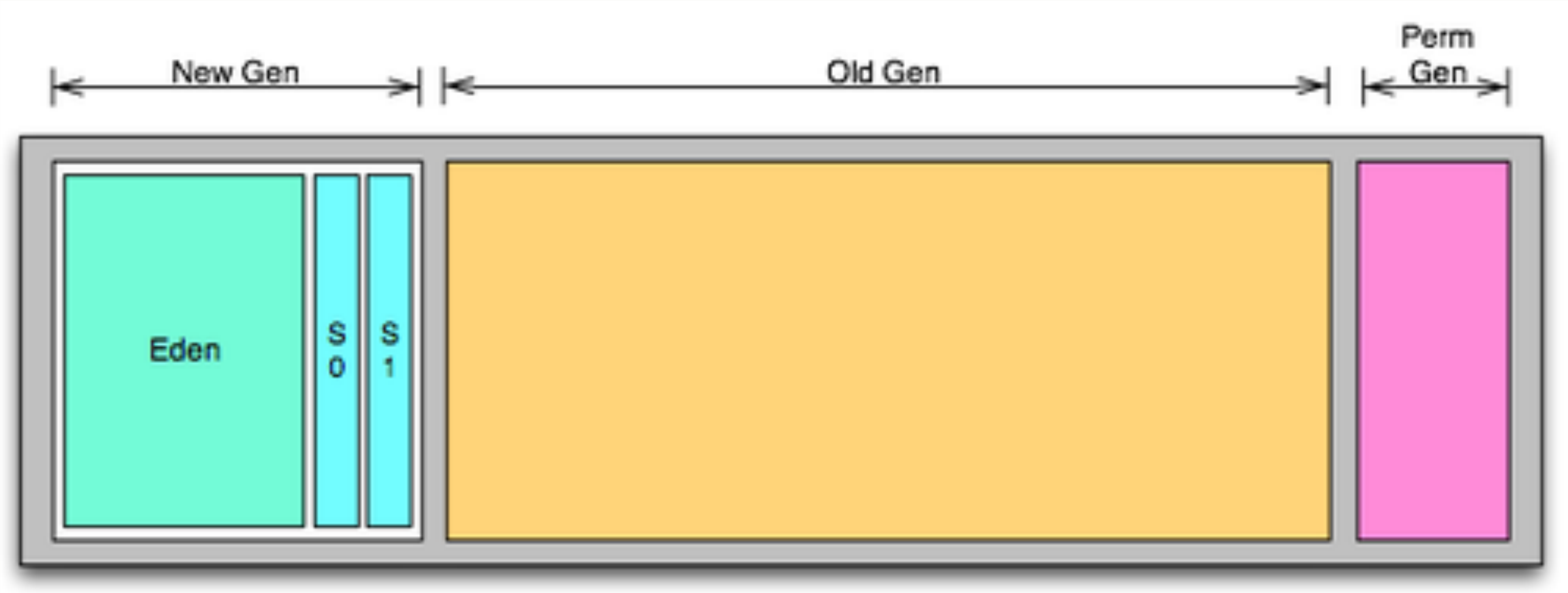


```
cqlsh:tutorial> TRACING on;
Now tracing requests.
cqlsh:tutorial> select * from tombstone_mayhem where pk=1 limit 100;

(0 rows)


Tracing session: 9a2039c0-33c3-11e4-93e5-05f76c346fb7

 activity                                                                  | timestamp    | source    | source_elapsed
---------------------------------------------------------------------------+--------------+-----------+----------------
                                                        execute_cql3_query | 16:39:52,541 | 127.0.0.1 |              0
             Parsing select * from tombstone_mayhem where pk=1 limit 100; | 16:39:52,541 | 127.0.0.1 |            587
                                                       Preparing statement | 16:39:52,542 | 127.0.0.1 |           1059
                     Executing single-partition query on tombstone_mayhem | 16:39:52,545 | 127.0.0.1 |           4830
                                              Acquiring sstable references | 16:39:52,545 | 127.0.0.1 |           4841
                                               Merging memtable tombstones | 16:39:52,546 | 127.0.0.1 |           4884
                      Partition index with 60 entries found for sstable 6 | 16:39:52,546 | 127.0.0.1 |           5704
                              Seeking to partition beginning in data file | 16:39:52,546 | 127.0.0.1 |           5714
                      Partition index with 24 entries found for sstable 5 | 16:39:52,547 | 127.0.0.1 |           6251
                              Seeking to partition beginning in data file | 16:39:52,547 | 127.0.0.1 |           6259
                      Partition index with 48 entries found for sstable 4 | 16:39:52,548 | 127.0.0.1 |           6904
                              Seeking to partition beginning in data file | 16:39:52,548 | 127.0.0.1 |           6912
 Skipped 0/3 non-slice-intersecting sstables, included 0 due to tombstones | 16:39:52,548 | 127.0.0.1 |           7112
                               Merging data from memtables and 3 sstables | 16:39:52,548 | 127.0.0.1 |           7134
                                  Read 0 live and 100000 tombstoned cells | 16:39:58,242 | 127.0.0.1 |        5701629
                                                          Request complete | 16:39:58,927 | 127.0.0.1 |        6386374
```

JVM Garbage Collection

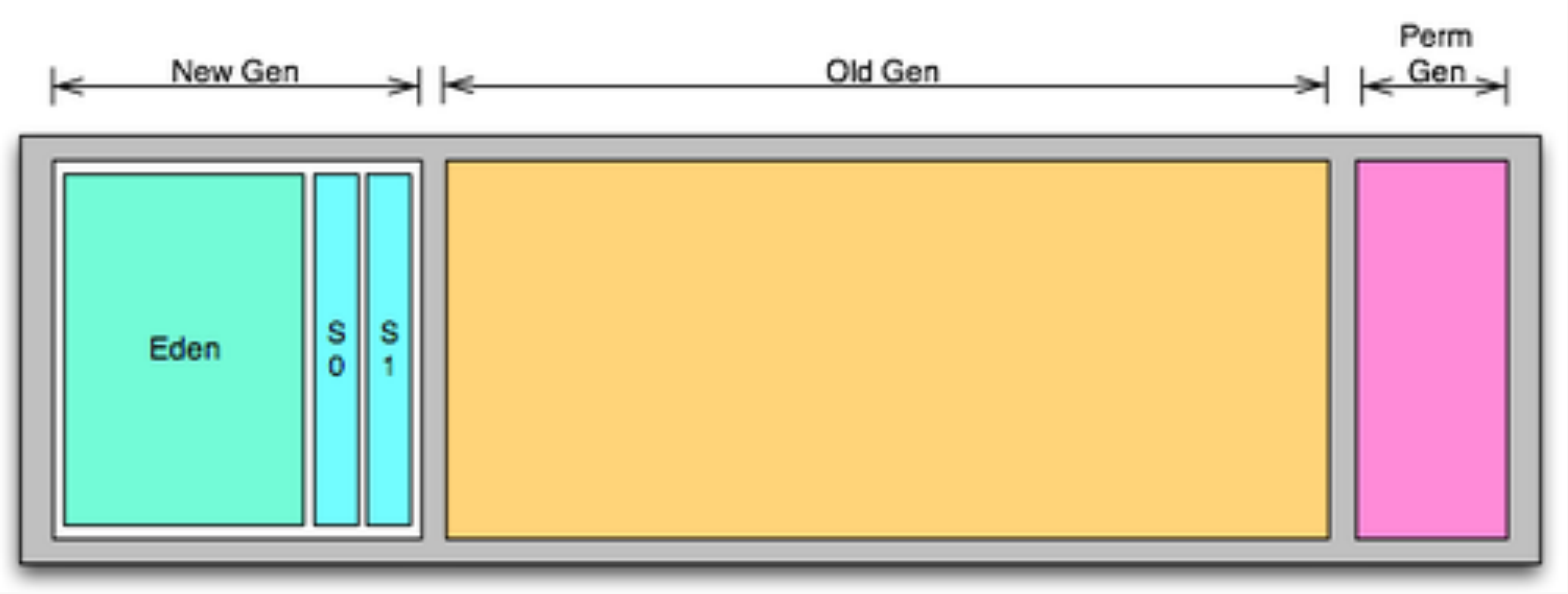
JVM GC Overview

- What is garbage collection?
  - Manual vs automatic memory management
- Generational garbage collection (ParNew & CMS)
  - New Generation
  - Old Generation



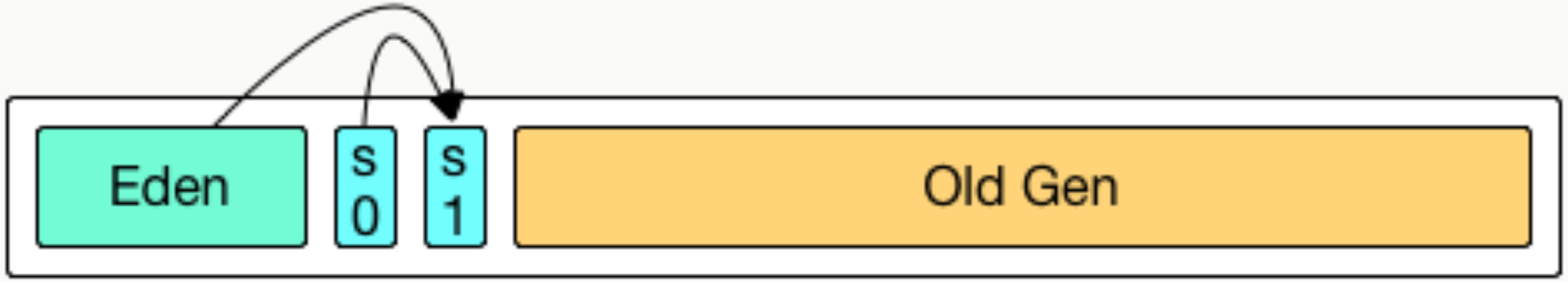
New Generation

- New objects are created in the new gen (eden)
- Comprised of Eden & 2 survivor spaces (SurvivorRatio)
- Space identified by HEAP_NEWSIZE in cassandra-env.sh
- Historically limited to 800MB



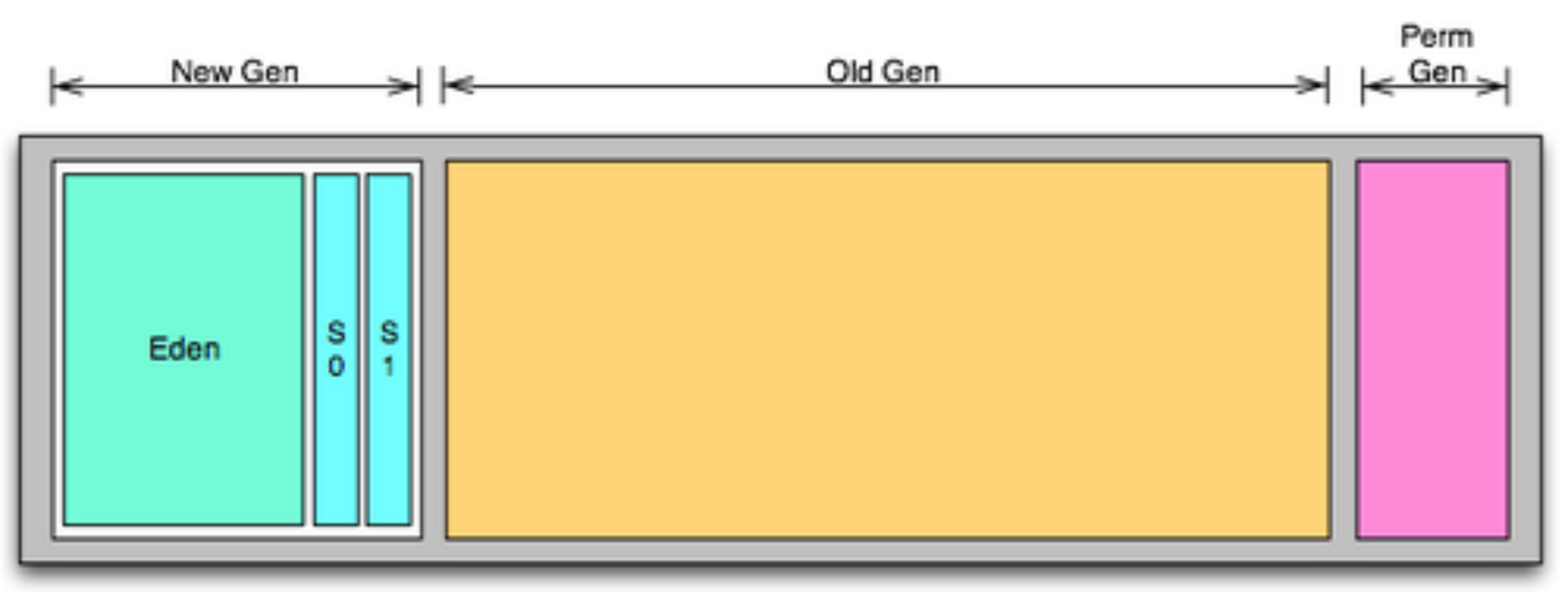
Minor GC

- Occurs when Eden fills up
- Stop the world
- Dead objects are removed
- Copy current survivor to empty survivor
- Live objects are promoted into survivor (S0 & S1) then old gen
- Some survivor objects promoted to old gen (MaxTenuringThreshold)
- Spillover promoted to old gen
- Removing objects is fast, promoting objects is slow



Old Generation

- Objects are promoted to new gen from old gen
- Major GC
  - Mostly concurrent
  - 2 short stop the world pauses

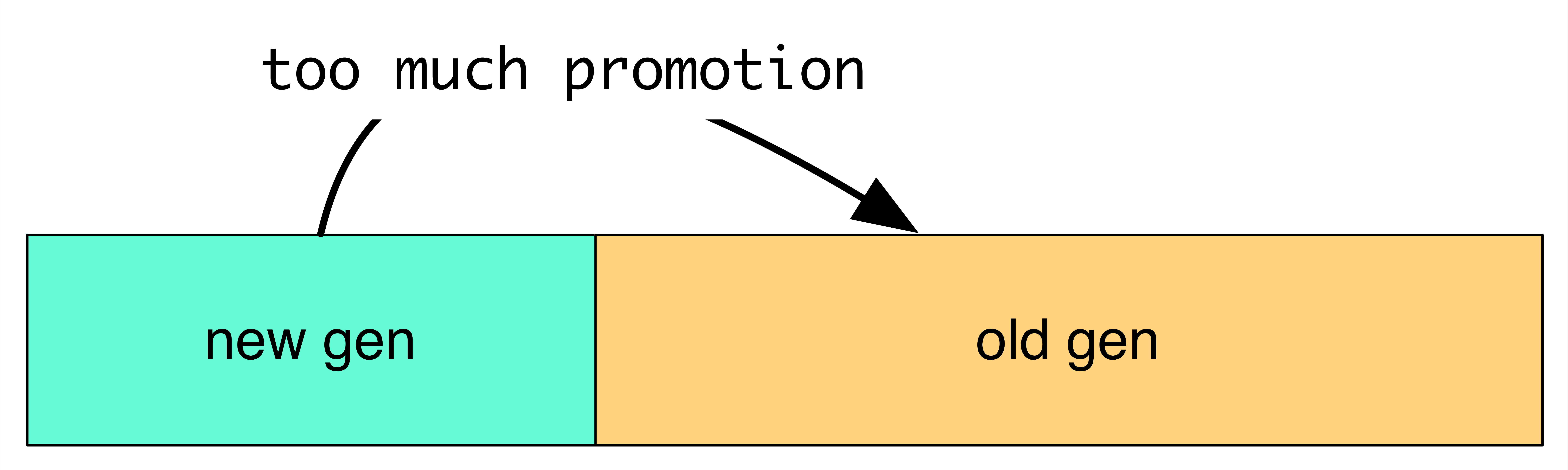


Full GC

- Occurs when old gen fills up or objects can’t be promoted
- Stop the world
- Collects all generations
- Defragments old gen
- These are bad!
- Massive pauses

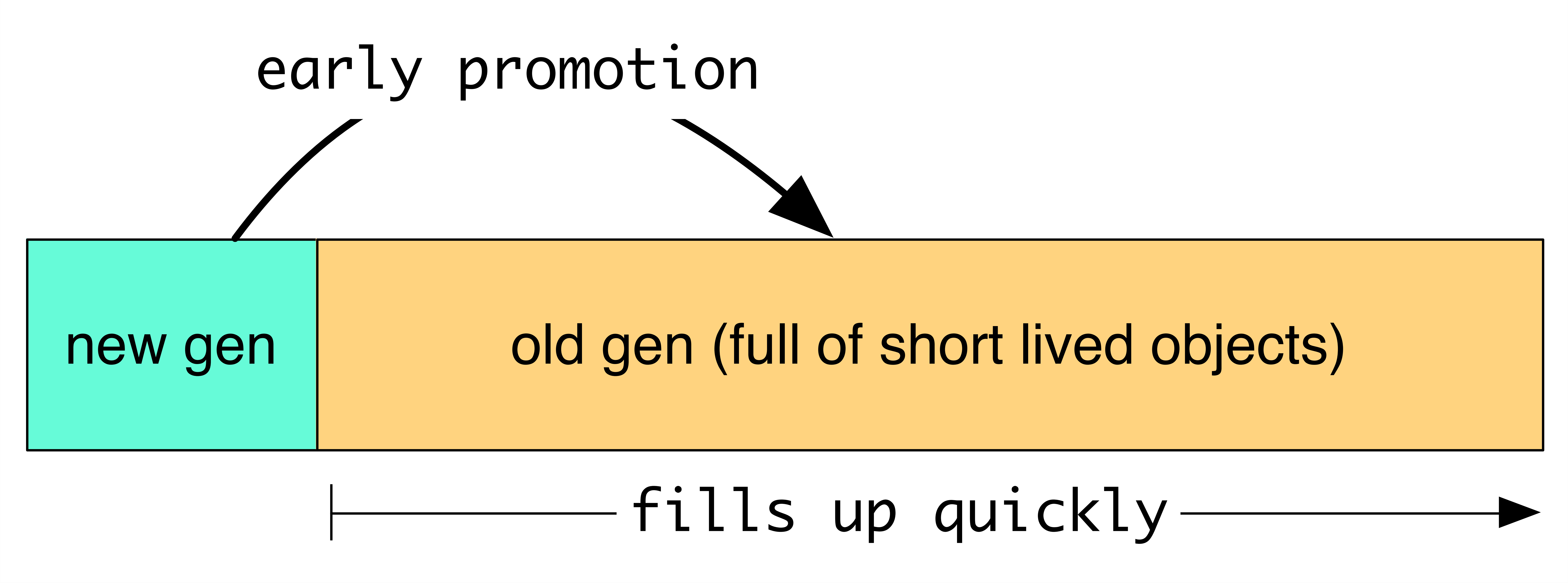
Workload 1: Write Heavy

- Objects promoted: Memtables
- New gen too big
- Remember: promoting objects is slow!
- Huge new gen = potentially a lot of promotion



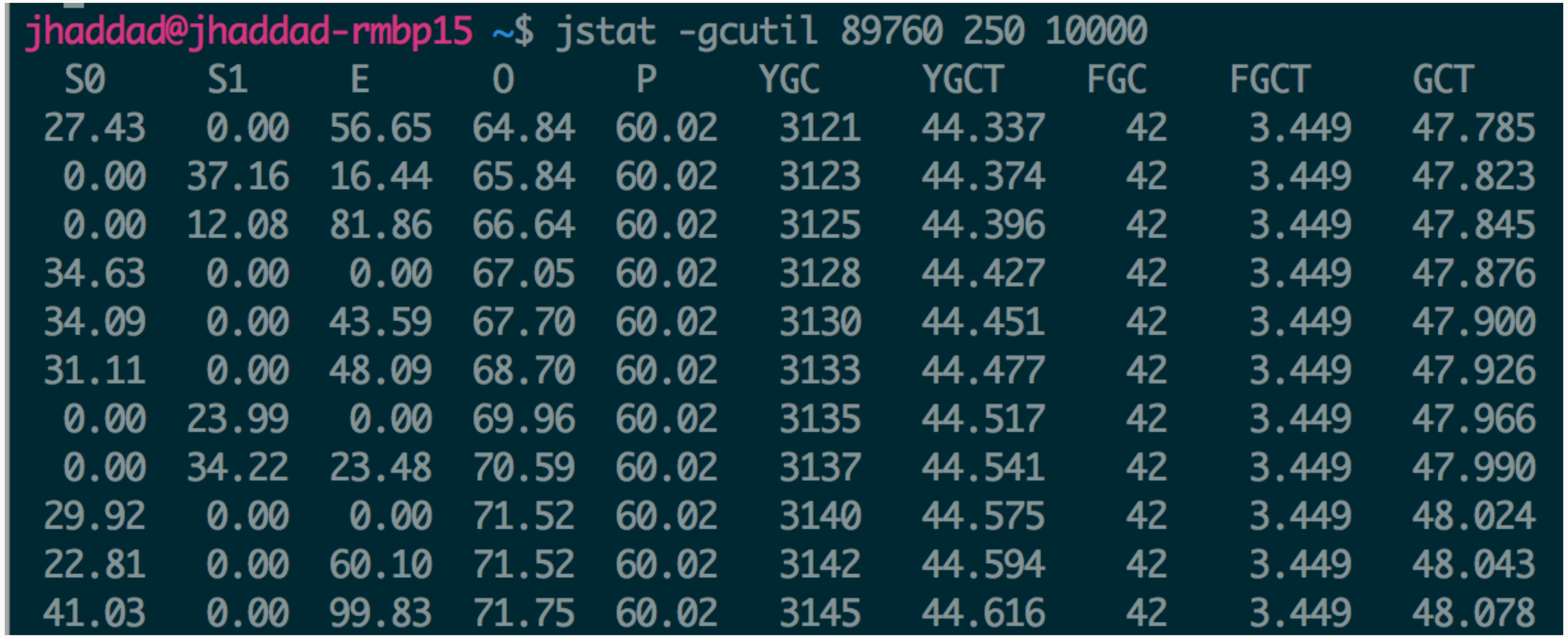
Workload 2: Read Heavy

- Short lived objects being promoted into old gen
- Lots of minor GCs
- Read heavy workloads on SSD
- Results in frequent full GC



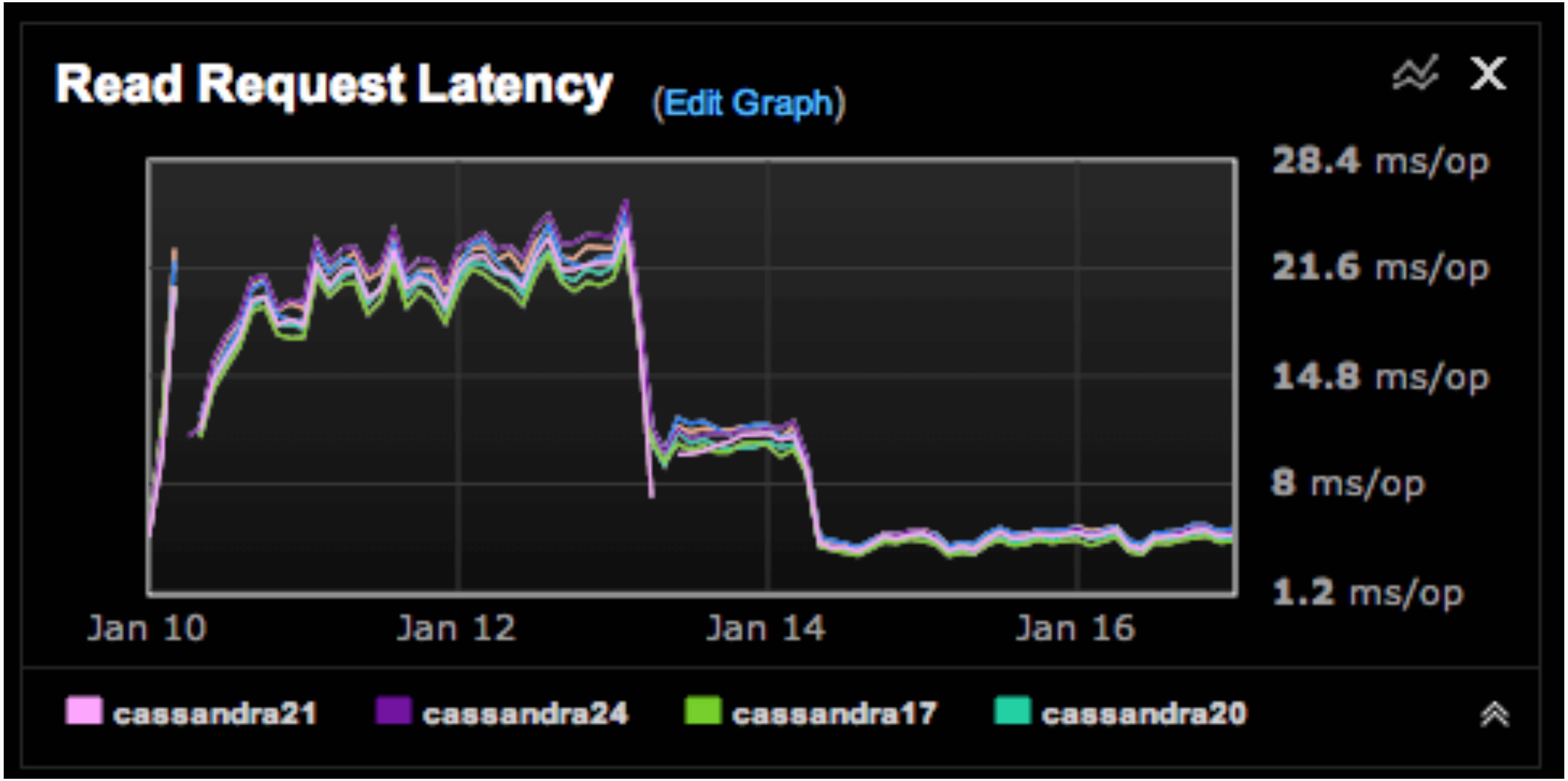
GC Profiling

- Opscenter gc stats
  - Look for correlations between gc spikes and read/write latency
- Cassandra GC Logging
  - Can be activated in cassandra-env.sh
- jstat
  - prints gc activity



```
jhaddad@jhaddad-rmbp15 ~$ jstat -gcutil 89760 250 10000
  S0     S1     E      O      P     YGC     YGCT    FGC    FGCT     GCT
 27.43   0.00  56.65  64.84  60.02   3121   44.337    42    3.449   47.785
  0.00  37.16  16.44  65.84  60.02   3123   44.374    42    3.449   47.823
  0.00  12.08  81.86  66.64  60.02   3125   44.396    42    3.449   47.845
 34.63   0.00   0.00  67.05  60.02   3128   44.427    42    3.449   47.876
 34.09   0.00  43.59  67.70  60.02   3130   44.451    42    3.449   47.900
 31.11   0.00  48.09  68.70  60.02   3133   44.477    42    3.449   47.926
  0.00  23.99   0.00  69.96  60.02   3135   44.517    42    3.449   47.966
  0.00  34.22  23.48  70.59  60.02   3137   44.541    42    3.449   47.990
 29.92   0.00   0.00  71.52  60.02   3140   44.575    42    3.449   48.024
 22.81   0.00  60.10  71.52  60.02   3142   44.594    42    3.449   48.043
 41.03   0.00  99.83  71.75  60.02   3145   44.616    42    3.449   48.078
```

How much does it matter?



Stuff is broken, fix it!

Narrow Down the Problem

DATASTAX

- Is it even Cassandra? Check your metrics!
- Nodes flapping / failing
  - Check ops center
  - Dig into system metrics
- Slow queries
  - Find your bottleneck
  - Check system stats
  - JVM GC
  - Compaction
  - Histograms
  - Tracing