Couchbase Performance and Scalability

Iterating with DTrace Observability

Dustin Sallings & Matt Ingenthron
@dlsspy @ingenthr

Who Are You?

Makers of a high performance document database.

Tell me More

Distributed across many systems or VMs, scale quickly, no special nodes

Simple to use, consistent by design

Managed cache, sysadmin managed resources, observability into operation of cluster

Who Uses It?

Ad targeting: submillisecond response time

Social Gaming: high throughput, scale fast

Social apps/content

COUCHBASE

About • Sign Out

MONITOR

- Cluster Overview
- Data Buckets
- Server Nodes
- Log

MANAGE

- Data Buckets
- Server Nodes
- Settings

SUPPORT

- Documentation
- Support Forums

Cluster Overview

Cluster

RAM Overview

Total Allocated (97.6 GB) — Total in Cluster (97.6 GB)

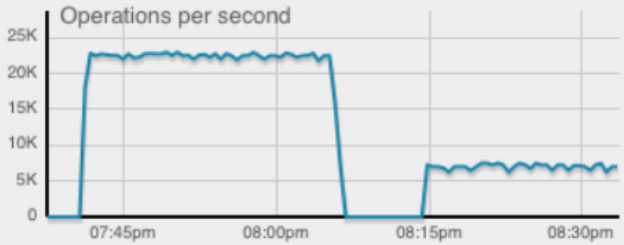
In Use (69.1 GB) — Unused (28.5 GB) — Unallocated (0 B)

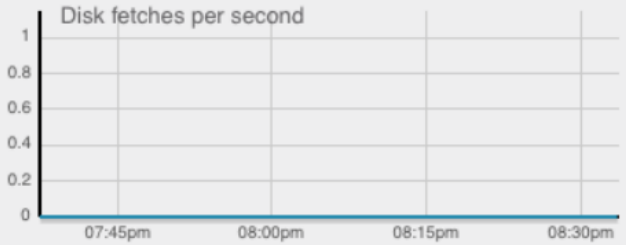
Disk Overview

Usable Free Space (777 GB) — Total Cluster Storage (984 GB)

In Use (192 GB) — Other Data (14 GB) — Free (777 GB)

Buckets (1 bucket active)





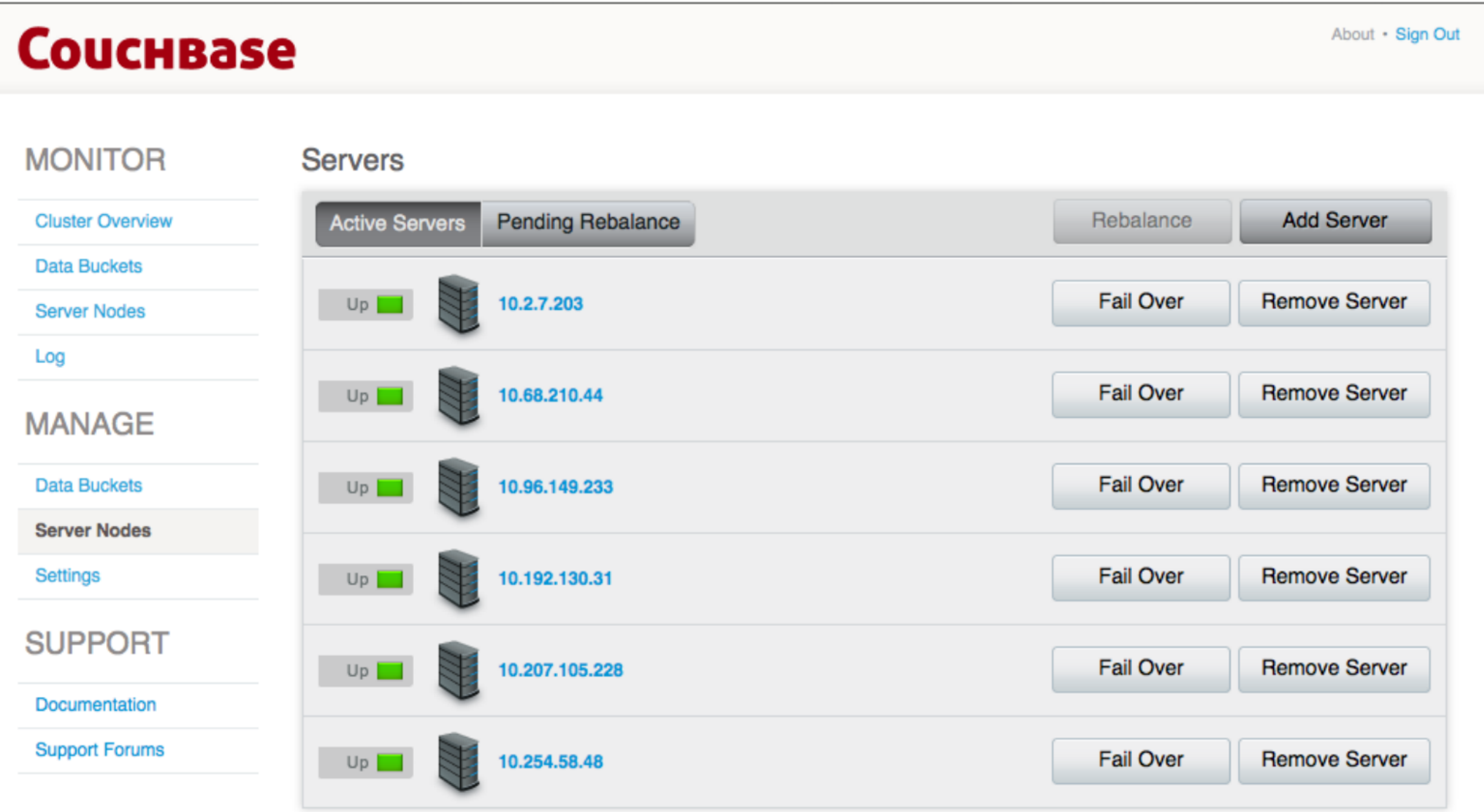
Servers

Active Servers: 10

Servers Failed Over: 0

Servers Down: 0

Servers Pending Rebalance: 0


Couchbase
About • Sign Out
MONITOR
Cluster Overview
Data Buckets
Server Nodes
Log
MANAGE
Data Buckets
Server Nodes
Settings
SUPPORT
Documentation
Support Forums
Servers
Active Servers
Pending Rebalance
Rebalance
Add Server
Up
10.2.7.203
Fail Over
Remove Server
Up
10.68.210.44
Fail Over
Remove Server
Up
10.96.149.233
Fail Over
Remove Server
Up
10.192.130.31
Fail Over
Remove Server
Up
10.207.105.228
Fail Over
Remove Server
Up
10.254.58.48
Fail Over
Remove Server

Couchbase

MONITOR

- Cluster Overview
- Data Buckets
- Server Nodes
- Log

MANAGE

- Data Buckets
- Server Nodes
- Settings

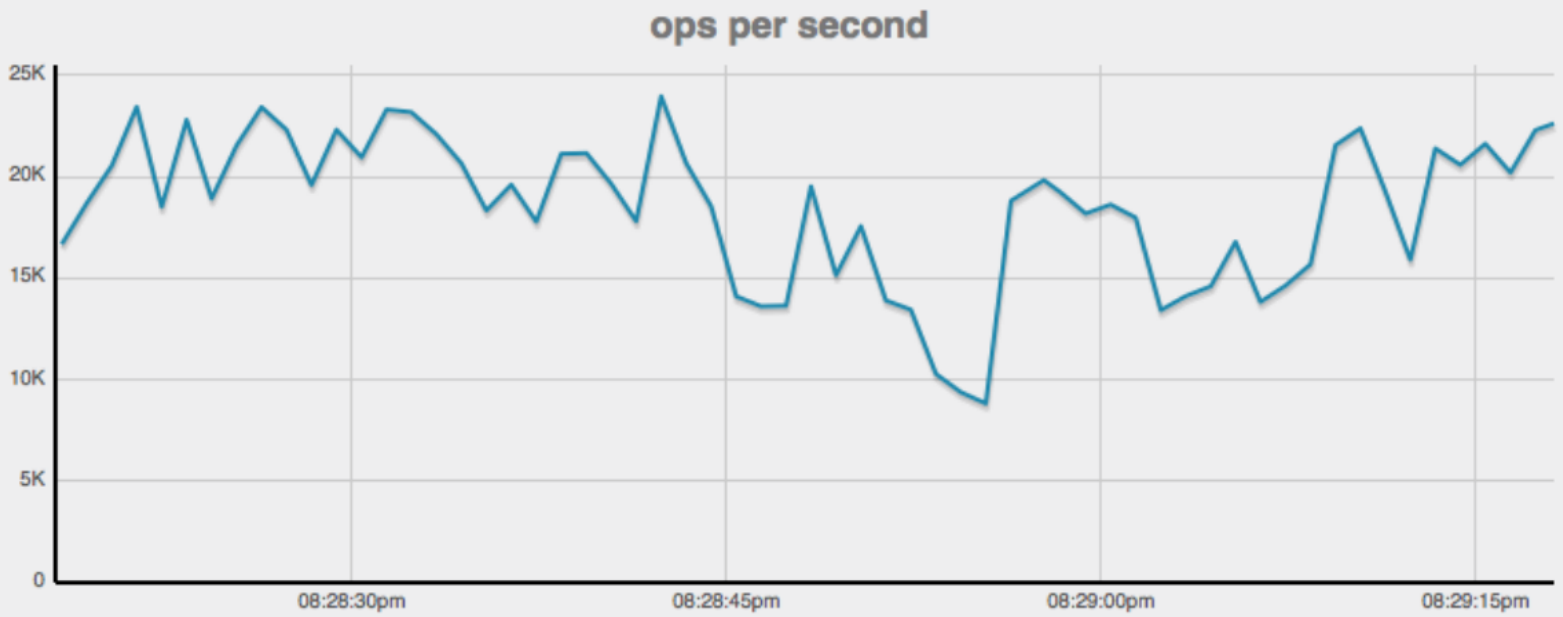
SUPPORT

- Documentation
- Support Forums

DATA BUCKETS: default on All Server Nodes

Minute | Hour | Day | Week | Month | Year

Last 1 minute

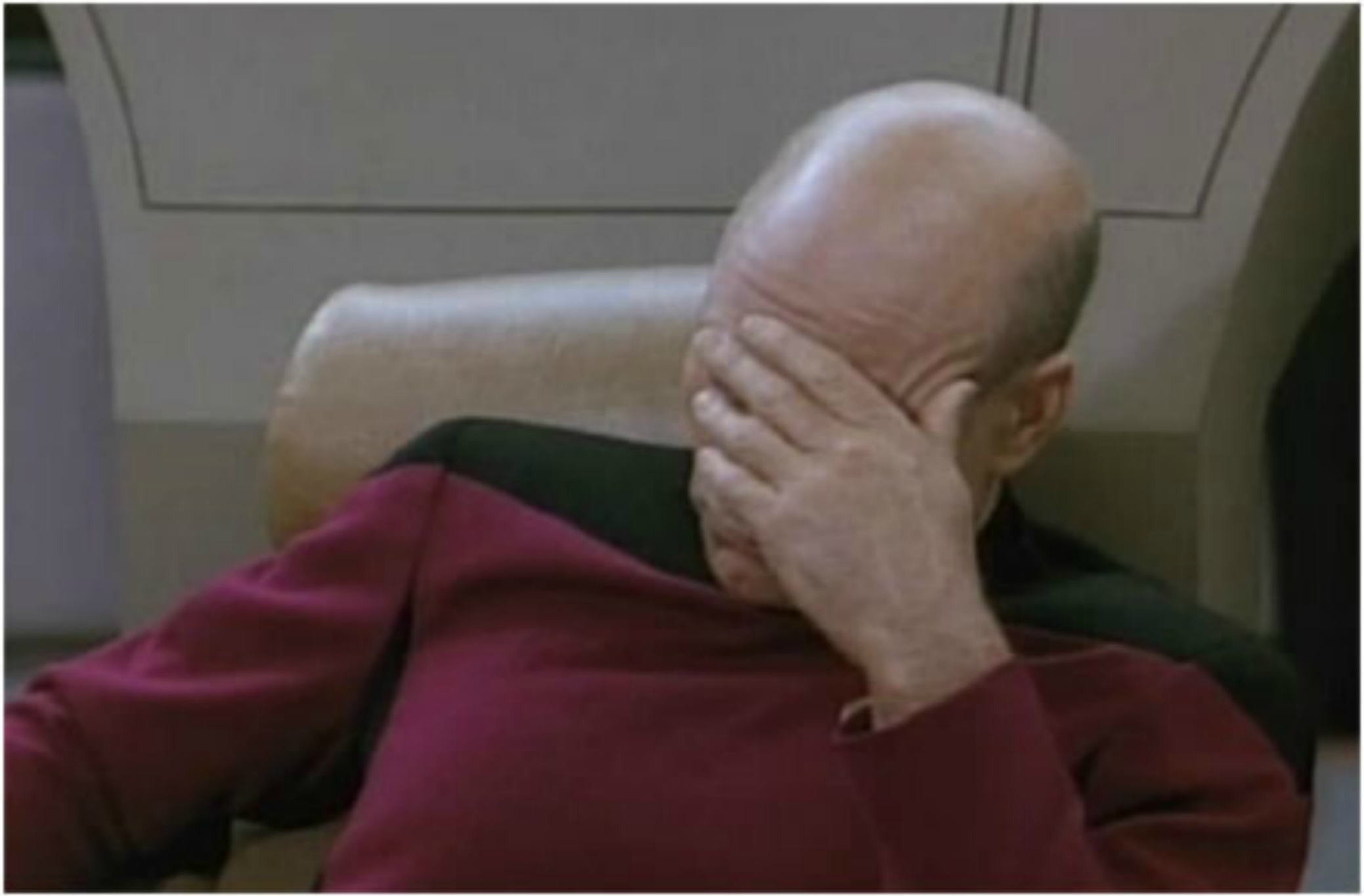


SUMMARY

22.7K ops per second	0.764 cache miss ratio	0 creates per sec.	5.61K updates per sec.
117 disk reads per sec.	0 temp OOM per sec.	15.3K gets per sec.	7.43K sets per sec.
0 CAS ops per sec.	0 deletes per sec.	10M items	5.01M disk write queue

Problems we ran into.

Sometimes thing are slow.



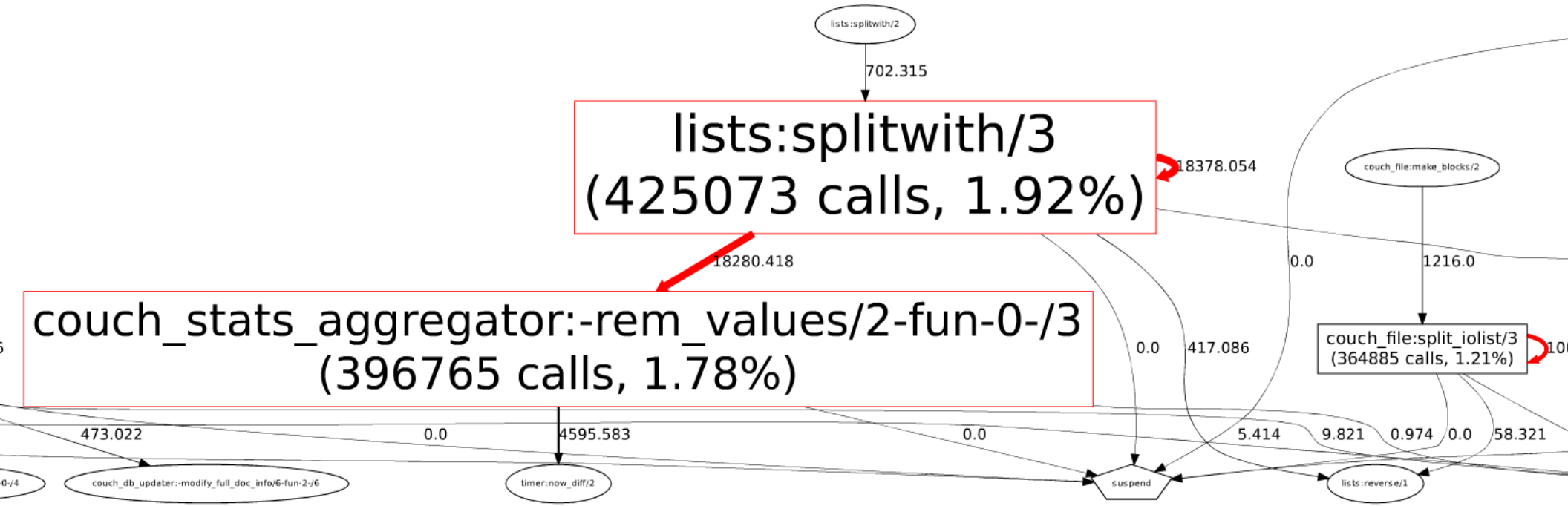
a poor craftsman blames his tools

Solve problems with data!

```
%                                                  CNT       ACC       OWN
[{ "<6499.675.0>",                                9564168,undefined,134393.587},    %%
 { spawned_by, "<6499.674.0>"},
 { spawned_as, {proc_lib,init_p,
                         ["<6499.674.0>",
                          ["<6499.668.0>"],
                          gen,init_it,
                          [gen_server,"<6499.674.0>","<6499.674.0>",
                           couch_db_updater,
                           {"<6499.674.0>",<<"default/0">>,
                            "/Users/dustin/Library/Application Support/CouchbaseServer/default/0.couch",
                            "<6499.669.0>",
                            [create]},
                           []]]}},
 { initial_calls, [{proc_lib,init_p,5},{erlang,put,2}]}].

{[{{gen_server,handle_msg,5},                     232,182645.632,    3.349}],
 { {gen_server,dispatch,3},                       232,182645.632,    3.349},     %
 [{{couch_db_updater,handle_info,2},              232,182642.283,   21.162}]}.

{[{{gen_server,dispatch,3},                       232,182642.283,   21.162}],
 { {couch_db_updater,handle_info,2},              232,182642.283,   21.162},     %
 [{{couch_db_updater,update_docs_int,5},          182,138567.114,   32.884},
  {{couch_db_updater,commit_data,1},               50,40266.369,    0.957},
  {{couch_db_updater,collect_updates,4},          181, 2188.293,    6.340},
  {{couch_db_updater,'-handle_info/2-lc$^0/1-0-',2}, 182, 1433.717,  360.113},
  {{gen_server,call,2},                           232,  145.618,    4.959},
  {{couch_db_updater,'-handle_info/2-lc$^3/1-3-',1}, 182,   10.710,    2.912},
  {{couch_db_update_notifier,notify,1},           181,    7.188,    2.166},
  {{couch_db_updater,'-handle_info/2-lc$^2/1-2-',2}, 182,    2.112,    2.111}]}.
```

60MB of this stuff

```
%                                                CNT       ACC       OWN
[{ "<6499.675.0>",                              9564168,undefined,134393.587},    %%
 { spawned_by, "<6499.674.0>"},
 { spawned_as, {proc_lib,init_p,
                         ["<6499.674.0>",
                          ["<6499.668.0>"],
                          gen,init_it,
                          [gen_server,"<6499.674.0>","<6499.674.0>",
                           couch_db_updater,
                           {"<6499.674.0>",<<"default/0">>},
                            "/Users/dustin/Library/Application Support/CouchbaseServer/default/0.couch",
                            "<6499.669.0>",
                            [create]},
                          []]]}},
 { initial_calls, [{proc_lib,init_p,5},{erlang,put,2}]}].

{[{{gen_server,handle_msg,5},                    232,182645.632,    3.349}],
 { {gen_server,dispatch,3},                      232,182645.632,    3.349},     %
 [{{couch_db_updater,handle_info,2},             232,182642.283,   21.162}]}.

{[{{gen_server,dispatch,3},                      232,182642.283,   21.162}],
 { {couch_db_updater,handle_info,2},             232,182642.283,   21.162},     %
 [{{couch_db_updater,update_docs_int,5},         182,138567.114,   32.884},
  {{couch_db_updater,commit_data,1},              50,40266.369,    0.957},
  {{couch_db_updater,collect_updates,4},         181, 2188.293,    6.340},
  {{couch_db_updater,'-handle_info/2-lc$^0/1-0-',2}, 182, 1433.717,  360.113},
  {{gen_server,call,2},                          232,  145.618,    4.959},
  {{couch_db_updater,'-handle_info/2-lc$^3/1-3-',1}, 182,   10.710,    2.912},
  {{couch_db_update_notifier,notify,1},          181,    7.188,    2.166},
  {{couch_db_updater,'-handle_info/2-lc$^2/1-2-',2}, 182,    2.112,    2.111}]}.
```


lists:splitwith/2
702.315
lists:splitwith/3
(425073 calls, 1.92%)
18378.054
couch_file:make_blocks/2
18280.418
0.0
1216.0
couch_stats_aggregator:-rem_values/2-fun-0-/3
(396765 calls, 1.78%)
0.0
417.086
couch_file:split_iolist/3
(364885 calls, 1.21%)
473.022
0.0
4595.583
0.0
5.414
9.821
0.974
0.0
58.321
couch_db_updater:-modify_full_doc_info/6-fun-2-/6
timer:now_diff/2
suspend
lists:reverse/1

Other tools?

yes, but...

tracing is harrrrrd

erlang:trace/3? percept?
ttb?, fprof?, eprof?, dbg?

erlang:trace/3

breaking the world?

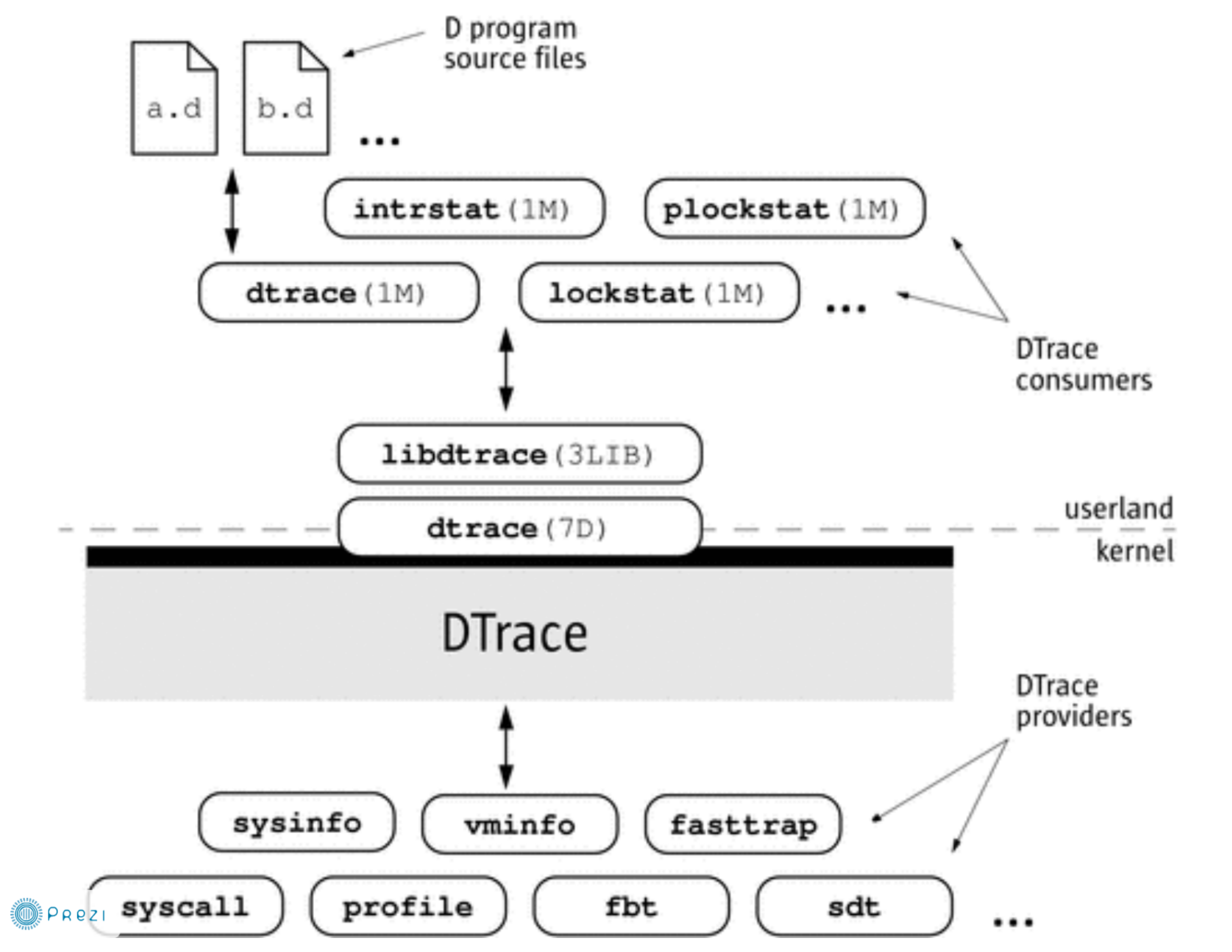
but I still don't know...

- what's leading to memory pressure
- scheduling events
- OS/device interactions
- internal messaging
- cross node messaging

enter: DTrace

enter: DTrace

...but we already have tracing!


D program source files
a.d
b.d
...
intrstat(1M)
plockstat(1M)
dtrace(1M)
lockstat(1M)
...
DTrace consumers
libdtrace(3LIB)
dtrace(7D)
userland
kernel
DTrace
DTrace providers
sysinfo
vminfo
fasttrap
syscall
profile
fbt
sdt
...

what's your point?

my program is more than what runs in the erlang vm

erlang dtrace gives me

- function, bif, nif entry and exit (separately, with pid, name, etc...)
- object copies (with sizes)
- process spawn and exit
- process scheduling events
- messages sent (with sizes)
- major and minor GC events (with sizes)

I'd like to also know...

I'd like to also know...

- what's triggering heap expansion
- what is causing actual IOPS
- how effectively does my app use the filesystem cache?
- And 239,446 other things I have probes for...

Where do I start?

Where do I start?

Usually not with DTrace.
Find a question.

Who's copying stuff?

```
 1 self string current;
 2
 3 erlang*:::function-entry,
 4 erlang*:::function-return,
 5 erlang*:::process_scheduled
 6 {
 7     self->current = copyinstr(arg1);
 8 }
 9
10 erlang*:::copy_struct
11 / self->current != "" /
12 {
13     @copies[self->current] = sum(arg0);
14 }
15
16 tick-10sec
17 {
18     trunc(@copies, 25);
19     printa(@copies);
20 }
```


dtrace: script 'copies.d' matched 7 probes
CPU ID FUNCTION:NAME
2 243372 :tick-10sec
gen_server:handle_msg/5 36
dict:on_bucket/3 48
couch_db:get_db_info/1 52
couch_db:handle_call/3 52
gen_server:cast/2 78
couch_api_wrap_httpc:stream_data_self/5 119
couch_replicator:handle_cast/2 120
couch_server:handle_call/3 152
couch_httpc_pool:handle_call/3 198
ibrowse_http_client:hexlist_to_integer/1 238
timer:handle_info/2 446
inet_parse:hex/2 991
timer:handle_call/3 1053
couch_task_status:handle_cast/2 2610
lists:keystore2/4 3613
ibrowse_http_client:parse_11_response/2 8190
gen:do_call/4 9009

How long does it take for a message to be received?

```
dtrace -qn 'erlang*:::send { sent[copyinstr(arg1)] = timestamp }
erlang*:::process_scheduled / sent[copyinstr(arg0)] / { @t =
quantize(timestamp - sent[copyinstr(arg0)]);
sent[copyinstr(arg0)] = 0;}'
```

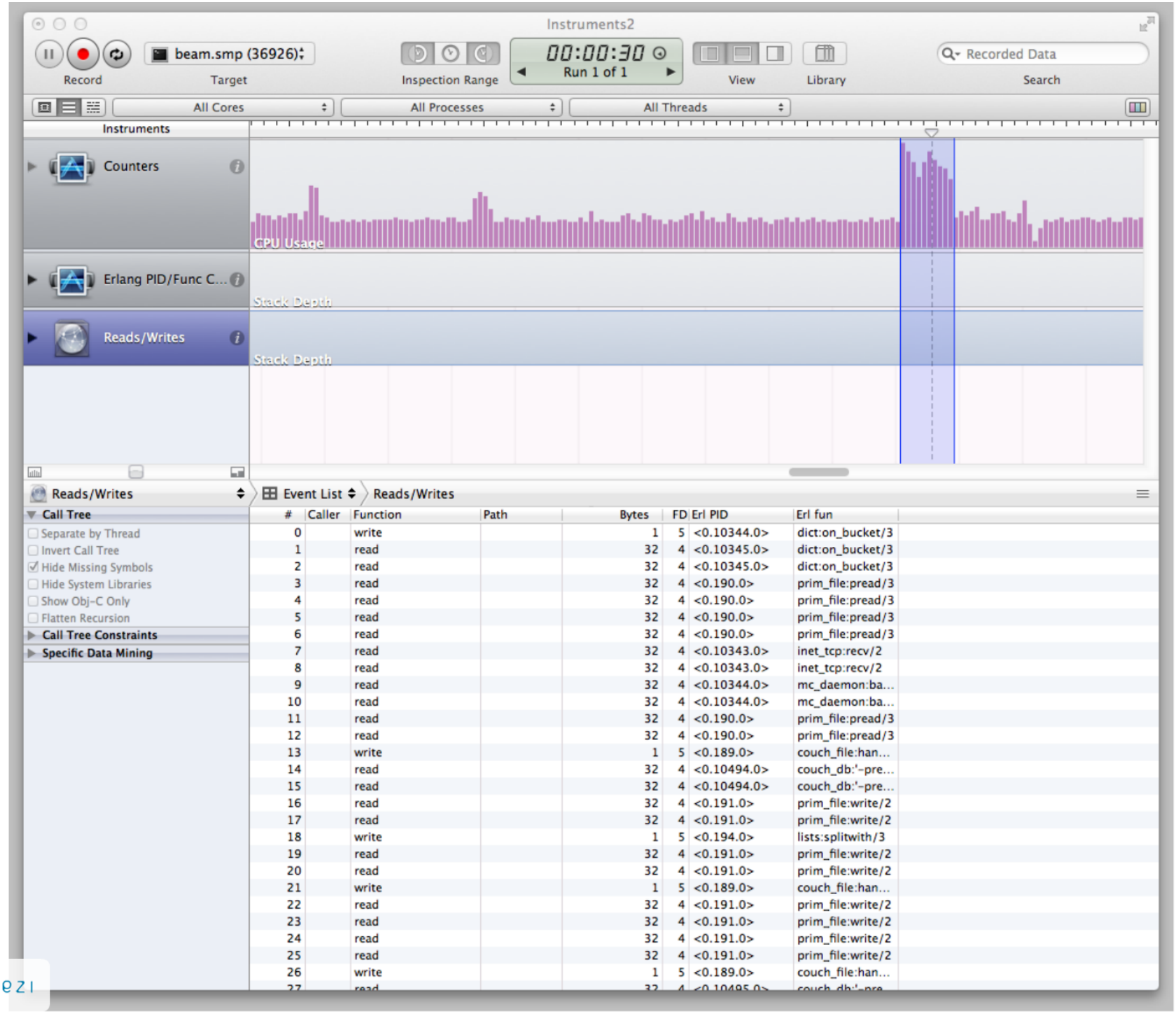
```
dtrace -qn 'erlang*:::send { sent[copyinstr(arg1)] = timestamp }
erlang*:::process_scheduled / sent[copyinstr(arg0)] / { @t =
quantize(timestamp - sent[copyinstr(arg0)]);
sent[copyinstr(arg0)] = 0;}'
```

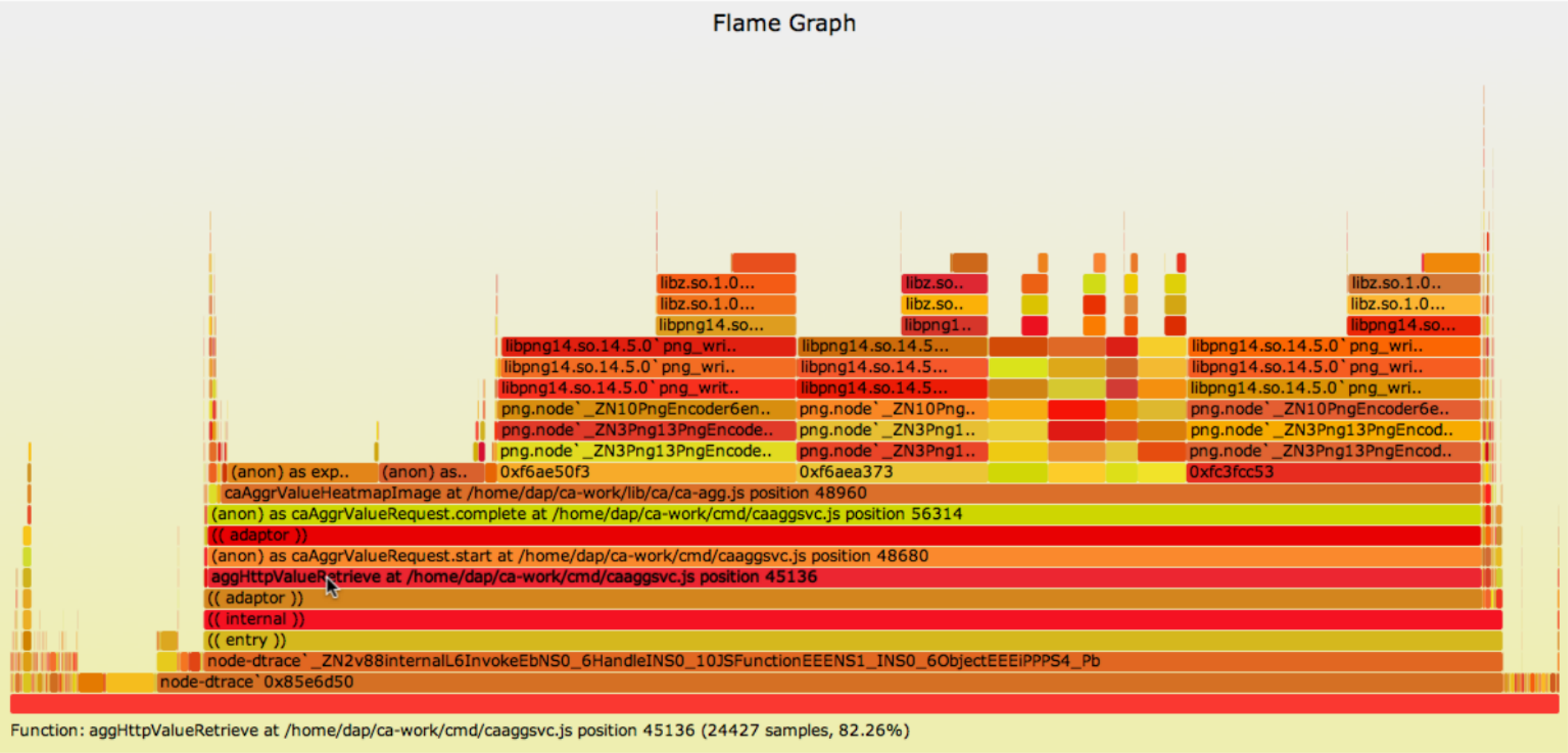
```
      value  ------------- Distribution ------------- count
      -1024 |                                         0
       -512 |                                         1
       -256 |                                         0
       -128 |                                         0
        -64 |                                         1
        -32 |                                         1
        -16 |                                         0
         -8 |                                         0
         -4 |                                         0
         -2 |                                         0
         -1 |                                         0
          0 |                                         0
          1 |                                         0
          2 |                                         0
          4 |                                         0
          8 |                                         0
         16 |                                         0
         32 |                                         0
         64 |                                         0
        128 |                                         2
        256 |                                         6
        512 |                                         8
       1024 |                                         26
       2048 |                                         66
       4096 |@@@@@                                    17049
       8192 |@@@@@@@@@@@@@@@@@@@                      58315
      16384 |@@@@@@@@@@                               31047
      32768 |@@@                                      10737
      65536 |@                                        3424
     131072 |@                                        1620
     262144 |                                         1469
     524288 |@                                        1614
    1048576 |@                                        2016
    2097152 |                                         151
    4194304 |                                         53
    8388608 |                                         47
   16777216 |                                         34
   33554432 |                                         44
   67108864 |                                         26
  134217728 |                                         28
  268435456 |                                         0
```

anything allocating currently?

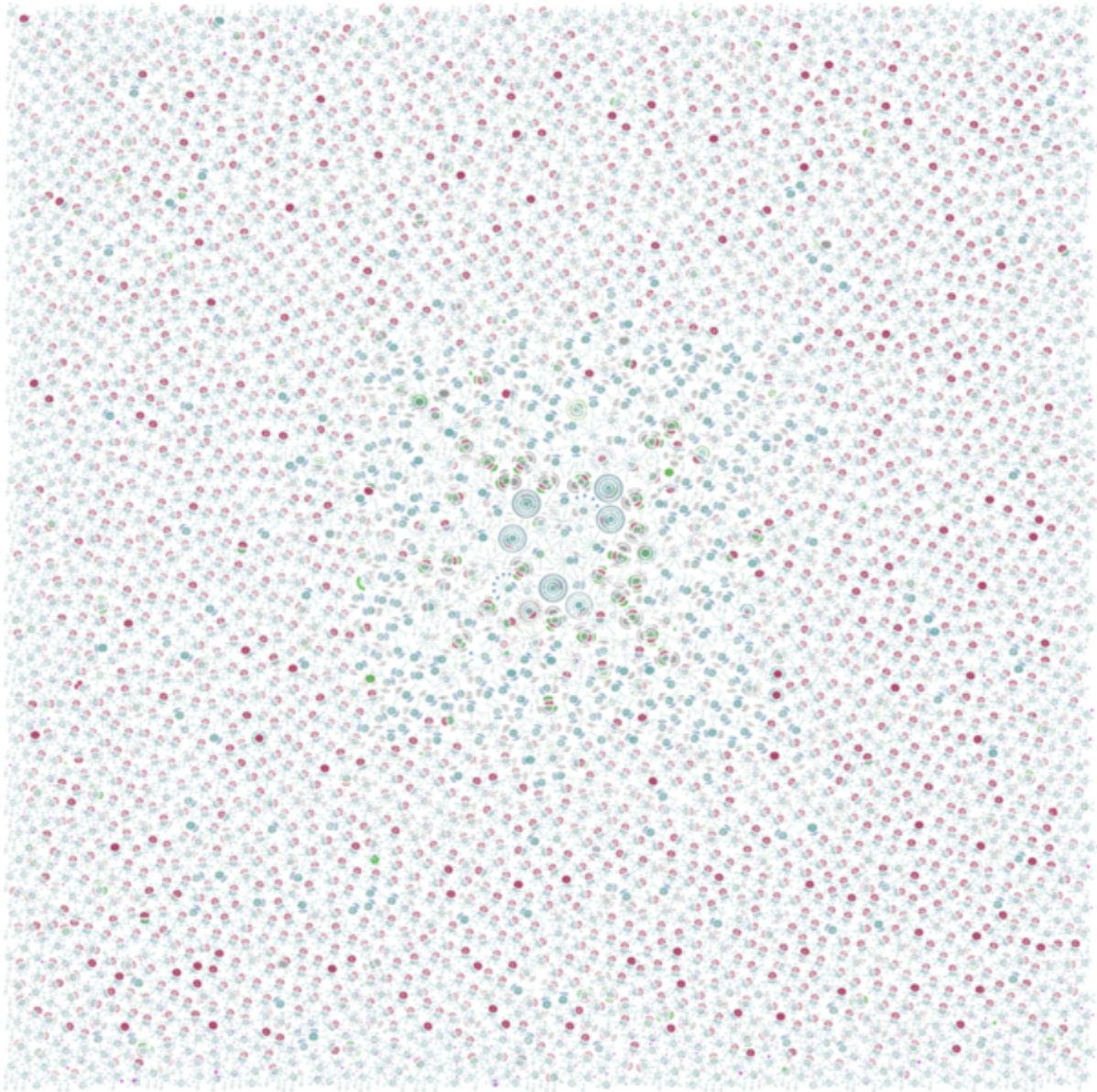
```
 1 self string current;
 2
 3 erlang*:::function-entry,
 4 erlang*:::function-return,
 5 erlang*:::process_scheduled
 6 / pid == $1 /
 7 {
 8     self->current = copyinstr(arg1);
 9 }
10
11 :::allocRegion
12 / pid == $1 && self->current != "" /
13 {
14     printf("Allocating near %s\n", self->current);
15 }
```

1 5813 szone_malloc_should_clear:allocRegion
Allocating near couch_api_wrap:receive_docs/4


Instruments2
Record
beam.smp (36926)
Target
Inspection Range
00:00:30
Run 1 of 1
View
Library
Recorded Data
Search
All Cores
All Processes
All Threads
Instruments
Counters
CPU Usage
Erlang PID/Func C...
Stack Depth
Reads/Writes
Stack Depth
Reads/Writes
Event List
Reads/Writes
Call Tree
Separate by Thread
Invert Call Tree
Hide Missing Symbols
Hide System Libraries
Show Obj-C Only
Flatten Recursion
Call Tree Constraints
Specific Data Mining
Caller Function Path Bytes FD Erl PID Erl fun
0 write 1 5 <0.10344.0> dict:on_bucket/3
1 read 32 4 <0.10345.0> dict:on_bucket/3
2 read 32 4 <0.10345.0> dict:on_bucket/3
3 read 32 4 <0.190.0> prim_file:pread/3
4 read 32 4 <0.190.0> prim_file:pread/3
5 read 32 4 <0.190.0> prim_file:pread/3
6 read 32 4 <0.190.0> prim_file:pread/3
7 read 32 4 <0.10343.0> inet_tcp:recv/2
8 read 32 4 <0.10343.0> inet_tcp:recv/2
9 read 32 4 <0.10344.0> mc_daemon:ba...
10 read 32 4 <0.10344.0> mc_daemon:ba...
11 read 32 4 <0.190.0> prim_file:pread/3
12 read 32 4 <0.190.0> prim_file:pread/3
13 write 1 5 <0.189.0> couch_file:han...
14 read 32 4 <0.10494.0> couch_db:'-pre...
15 read 32 4 <0.10494.0> couch_db:'-pre...
16 read 32 4 <0.191.0> prim_file:write/2
17 read 32 4 <0.191.0> prim_file:write/2
18 write 1 5 <0.194.0> lists:splitwith/3
19 read 32 4 <0.191.0> prim_file:write/2
20 read 32 4 <0.191.0> prim_file:write/2
21 write 1 5 <0.189.0> couch_file:han...
22 read 32 4 <0.191.0> prim_file:write/2
23 read 32 4 <0.191.0> prim_file:write/2
24 read 32 4 <0.191.0> prim_file:write/2
25 read 32 4 <0.191.0> prim_file:write/2
26 write 1 5 <0.189.0> couch_file:han...


Flame Graph
libz.so.1.0...
libz.so..
libz.so.1.0..
libpng14.so...
libpng1..
libpng14.so.14.5.0`png_wri..
libpng14.so.14.5...
png.node`_ZN10PngEncoder6en..
png.node`_ZN10Png..
png.node`_ZN3Png13PngEncode..
png.node`_ZN3Png1..
png.node`_ZN3Png13PngEncod..
(anon) as exp..
(anon) as..
0xf6ae50f3
0xf6aea373
0xfc3fcc53
caAggrValueHeatmapImage at /home/dap/ca-work/lib/ca/ca-agg.js position 48960
(anon) as caAggrValueRequest.complete at /home/dap/ca-work/cmd/caaggsvc.js position 56314
((adaptor))
(anon) as caAggrValueRequest.start at /home/dap/ca-work/cmd/caaggsvc.js position 48680
aggHttpValueRetrieve at /home/dap/ca-work/cmd/caaggsvc.js position 45136
((internal))
((entry))
node-dtrace`_ZN2v88internalL6InvokeEbNS0_6HandleINS0_10JSFunctionEEENS1_INS0_6ObjectEEEiPPPS4_Pb
node-dtrace`0x85e6d50
Function: aggHttpValueRetrieve at /home/dap/ca-work/cmd/caaggsvc.js position 45136 (24427 samples, 82.26%)

I'm Sold! Let's DTrace Everything!



oh, linux

"This is Not DTrace"

```
[root@screven ~]# uname -a
Linux screven 2.6.32-201.0.4.el6uek.x86_64 #1 SMP Tue Oct 4 16:47:00 EDT 2011 x86_64 x86_64 x86_
[root@screven ~]# dtrace -n 'BEGIN{ trace("howdy from linux"); }'
dtrace: description 'BEGIN' matched 1 probe
CPU     ID                    FUNCTION:NAME
0       1                           :BEGIN   howdy from linux
^C
```

Then I wanted to see what was on the system:

```
[root@screven ~]# dtrace -l | wc -l
574
```

574 < 239,446

http://dtrace.org/blogs/ahl/2011/10/10/oel-this-is-not-dtrace/

systemtap?

API compatible with DTrace
for probe portability, but...

...go ahead and try it somewhere

np, I run OS X, FreeBSD, SmartOS, Illumos, etc...

(and then deploy to Linux)

well, *almost* there, but the probes are broken in otp's master branch...

Join in the Tracing

Let us and others know if you are interested: "#dtrace for #erlang"

Let's build more regularly, keep it working (and prevent forks)

'mo betta probing. Function end? Ideas?

thanks to

@brendangregg
@slfritchie
@ahl

you

probing questions?

@ingenthr
@dlsspy