Building OpenDNS Stats

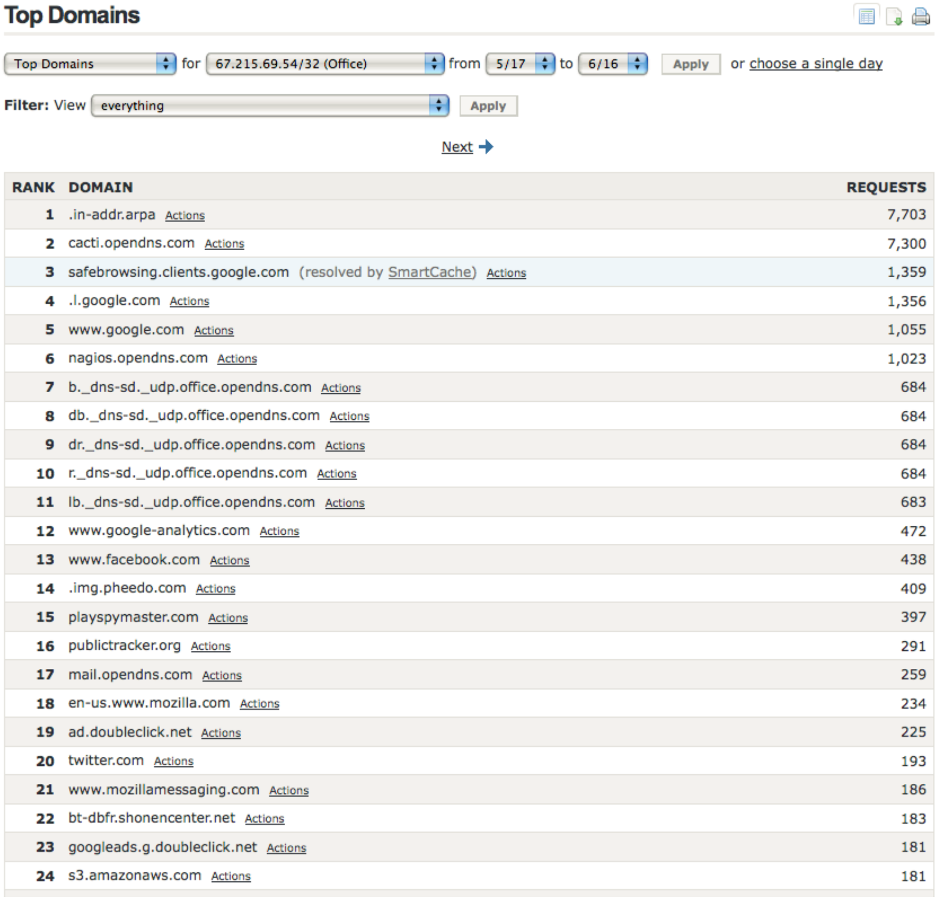
Richard Crowley
richard@opendns.com

Then: 8 billion DNS queries per day

```
                                                                          in.l.google.com. 1 0
                                                                          i.co.jp. 1 0
                                                                          . 1 0
@400000004a381ba80dd3ae34 q1 24.155.125.240 normal 1045953 my-iqquiz.com. 1 0
@400000004a381ba80dd3b604 q1 64.253.103.18 normal 788290 6.164.133.166.in-addr.arpa. 12 2
@400000004a381ba80dd3bdd4 q1 70.246.80.10 normal 0 googleads.g.doubleclick.net. 1 0
@400000004a381ba80dd3c5a4 q1 98.108.66.45 normal 0 _ldap._tcp.nj-bloomfield._sites.dc._msdcs.mrii.c
@400000004a381ba80dd41b94 q1 98.144.16.195 normal 0 js.casalemedia.com. 1 0
@400000004a381ba80dd42364 q1 68.165.29.60 normal 0 img-cdn.mediaplex.com. 1 0
@400000004a381ba80dd42b34 q1 12.233.75.219 normal 0 zsmseno.clnet.cz. 1 0
@400000004a381ba80dd43304 q1 174.37.58.88 normal 0 70.96.118.85.bl.spamcop.net. 16 0
@400000004a381ba80dd43ad4 q1 208.76.86.13 normal 519070 252.76.75.208.bl.spamcop.net. 1 3
@400000004a381ba80dd442a4 q1 201.138.19.196 normal 0 isatap.domain.local. 1 3
@400000004a381ba80dd465cc q1 24.192.98.53 normal 0 208.85.224.82.in-addr.arpa. 12 0
@400000004a381ba80dd46d9c q1 64.91.71.57 normal 0 liveupdate.symantecliveupdate.com. 1 0
@400000004a381ba80dd4756c q1 69.64.43.245 normal 558867 alt4.gmail-smtp-in.l.google.com. 1 0
@400000004a381ba80dd47d3c q1 69.64.43.245 normal 558867 alt4.gmail-smtp-in.l.google.com. 1 0
@400000004a381ba80dd4850c q1 72.10.191.11 normal 812477 iprep1.t.ctmail.com. 1 0
@400000004a381ba80dd49c7c q1 12.233.75.219 normal 0 zsmseno.clnet.cz. 1 0
@400000004a381ba80dd4a44c q1 69.157.60.79 normal 0 img-cdn.mediaplex.com. 1 0
@400000004a381ba80dd4ac1c q1 208.43.52.205 nxdomain 0 haghway.com.br. 1 0
@400000004a381ba80dd4b3ec q1 204.145.0.242 normal 488877 105.12.90.201.asetnhap5duax9a26l24rda5g3gv
@400000004a381ba80dd4bbbc q1 206.246.157.1 normal 0 penninegas.co.uk. 15 2
@400000004a381ba80dd4c38c q1 69.21.243.131 normal 0 svn.atomicobject.com. 28 0
@400000004a381ba80dd4dafc q1 163.192.13.65 normal 894966 dns.hitachi-koki.co.jp. 1 0
@400000004a381ba80dd4e2cc q1 76.65.199.42 nxdomain 0 cs16.msg.dcn.yahoo.com. 1 0
@400000004a381ba80dd4ea9c q1 189.169.97.227 normal 0 impaktosoo.gateway.2wire.net. 1 3
@400000004a381ba80dd4f26c q1 69.64.43.245 normal 558867 gmail.com. 15 0
@400000004a381ba80dd4f654 q1 189.168.174.182 normal 0 wpad.2wire.net. 1 3
@400000004a381ba80dd4fe24 q1 69.64.43.245 normal 558867 alt3.gmail-smtp-in.l.google.com. 1 0
@400000004a381ba80dd51594 q1 189.133.170.67 normal 0 v13.lscache5.googlevideo.com. 1 0
@400000004a381ba80dd538bc q1 12.186.60.189 nxdomain 0 carolyn5.ktemca.com. 1 0
@400000004a381ba80dd5408c q1 72.249.148.132 normal 384918 mailin-04.mx.aol.com. 1 0
@400000004a381ba80dd5485c q1 76.65.199.42 nxdomain 0 csa.yahoo.com. 1 0
@400000004a381ba80dd5502c
@400000004a381ba80dd55414
@400000004a381ba80dd55be4
```

Now: 14 billion DNS queries per day

Logs are silly, let's make graphs



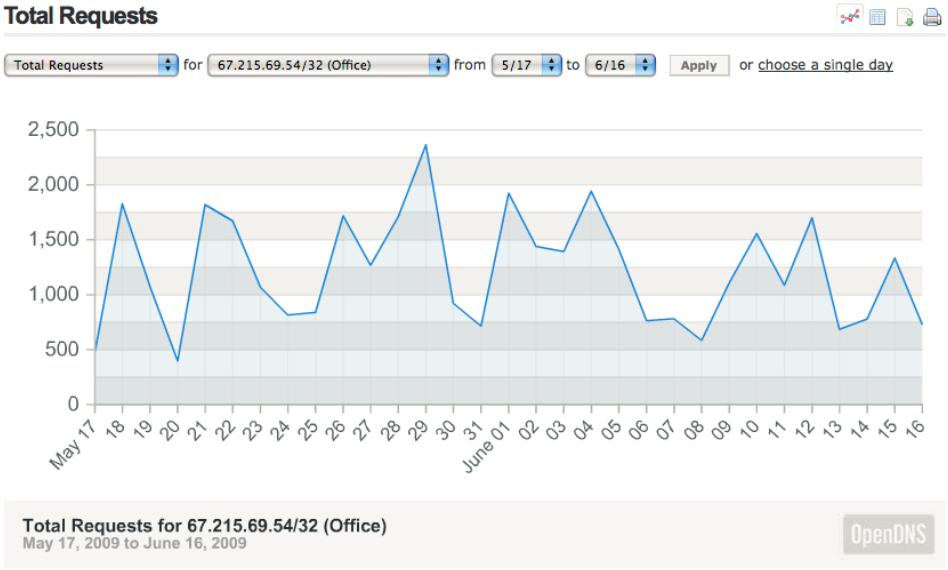
Top Domains

Top Domains ⬍ for 67.215.69.54/32 (Office) ⬍ from 5/17 ⬍ to 6/16 ⬍ Apply or choose a single day

Filter: View everything ⬍ Apply

Next →

RANK	DOMAIN	REQUESTS
1	.in-addr.arpa Actions	7,703
2	cacti.opendns.com Actions	7,300
3	safebrowsing.clients.google.com (resolved by SmartCache) Actions	1,359
4	.l.google.com Actions	1,356
5	www.google.com Actions	1,055
6	nagios.opendns.com Actions	1,023
7	b._dns-sd._udp.office.opendns.com Actions	684
8	db._dns-sd._udp.office.opendns.com Actions	684
9	dr._dns-sd._udp.office.opendns.com Actions	684
10	r._dns-sd._udp.office.opendns.com Actions	684
11	lb._dns-sd._udp.office.opendns.com Actions	683
12	www.google-analytics.com Actions	472
13	www.facebook.com Actions	438
14	.img.pheedo.com Actions	409
15	playspymaster.com Actions	397
16	publictracker.org Actions	291
17	mail.opendns.com Actions	259
18	en-us.www.mozilla.com Actions	234
19	ad.doubleclick.net Actions	225
20	twitter.com Actions	193
21	www.mozillamessaging.com Actions	186
22	bt-dbfr.shonencenter.net Actions	183
23	googleads.g.doubleclick.net Actions	181
24	s3.amazonaws.com Actions	181



Total Requests

Total Requests ⬍ for 67.215.69.54/32 (Office) ⬍ from 5/17 ⬍ to 6/16 ⬍ Apply or choose a single day

Total Requests for 67.215.69.54/32 (Office)
May 17, 2009 to June 16, 2009

High level design from my OpenDNS interview

map/reduce/ish

Stage 1 buckets data by network

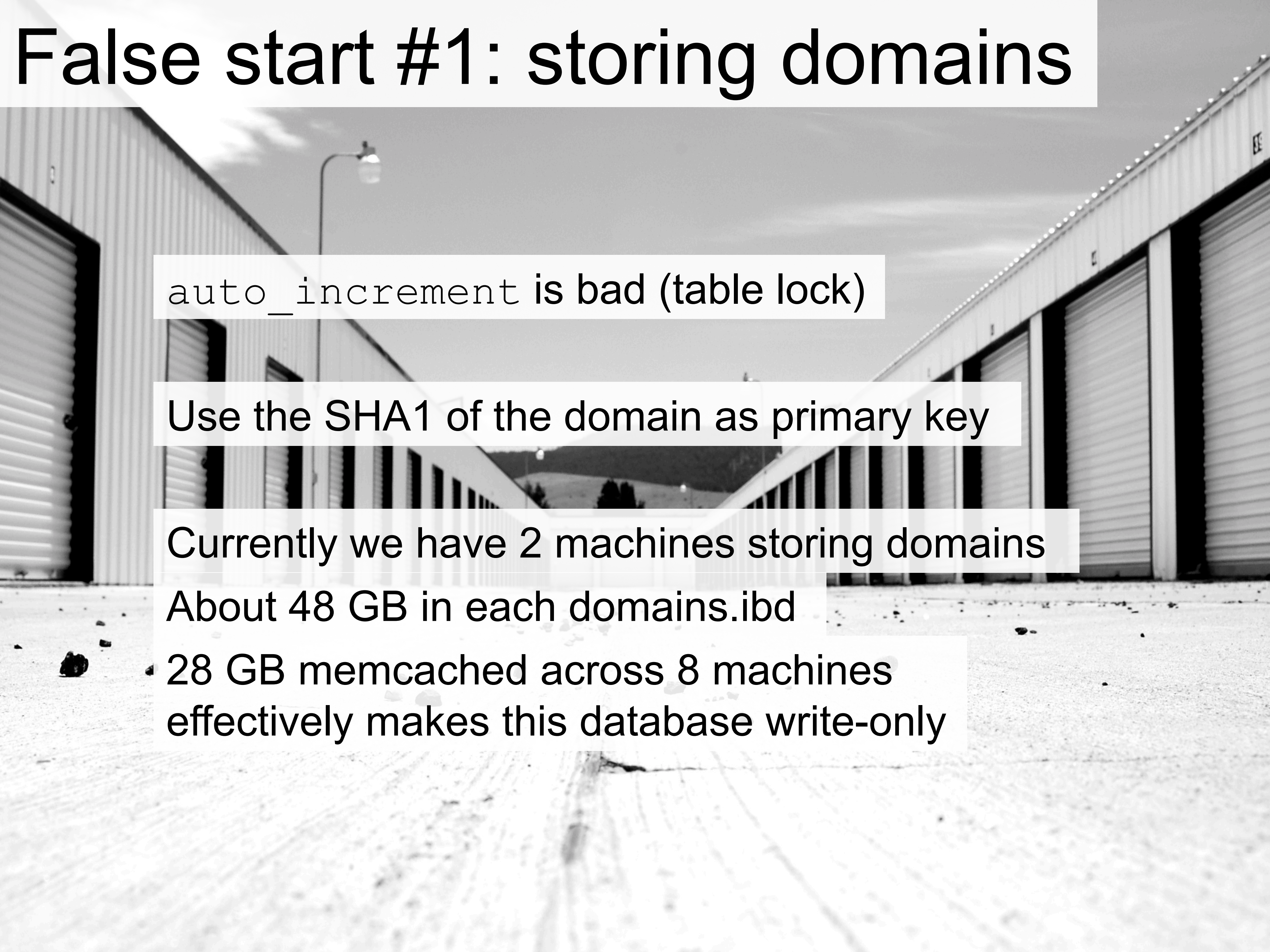
Stage 2 *aggregates* and stores

Prefers to duplicate data rather than omit data

Give each network a separate table (keeps each table small(er) and keeps the primary key small(er))

False starts

False start #1: storing domains



`auto_increment` is bad (table lock)

Use the SHA1 of the domain as primary key

Currently we have 2 machines storing domains

About 48 GB in each domains.ibd

28 GB memcached across 8 machines effectively makes this database write-only

False start #2: `std::bad_alloc`

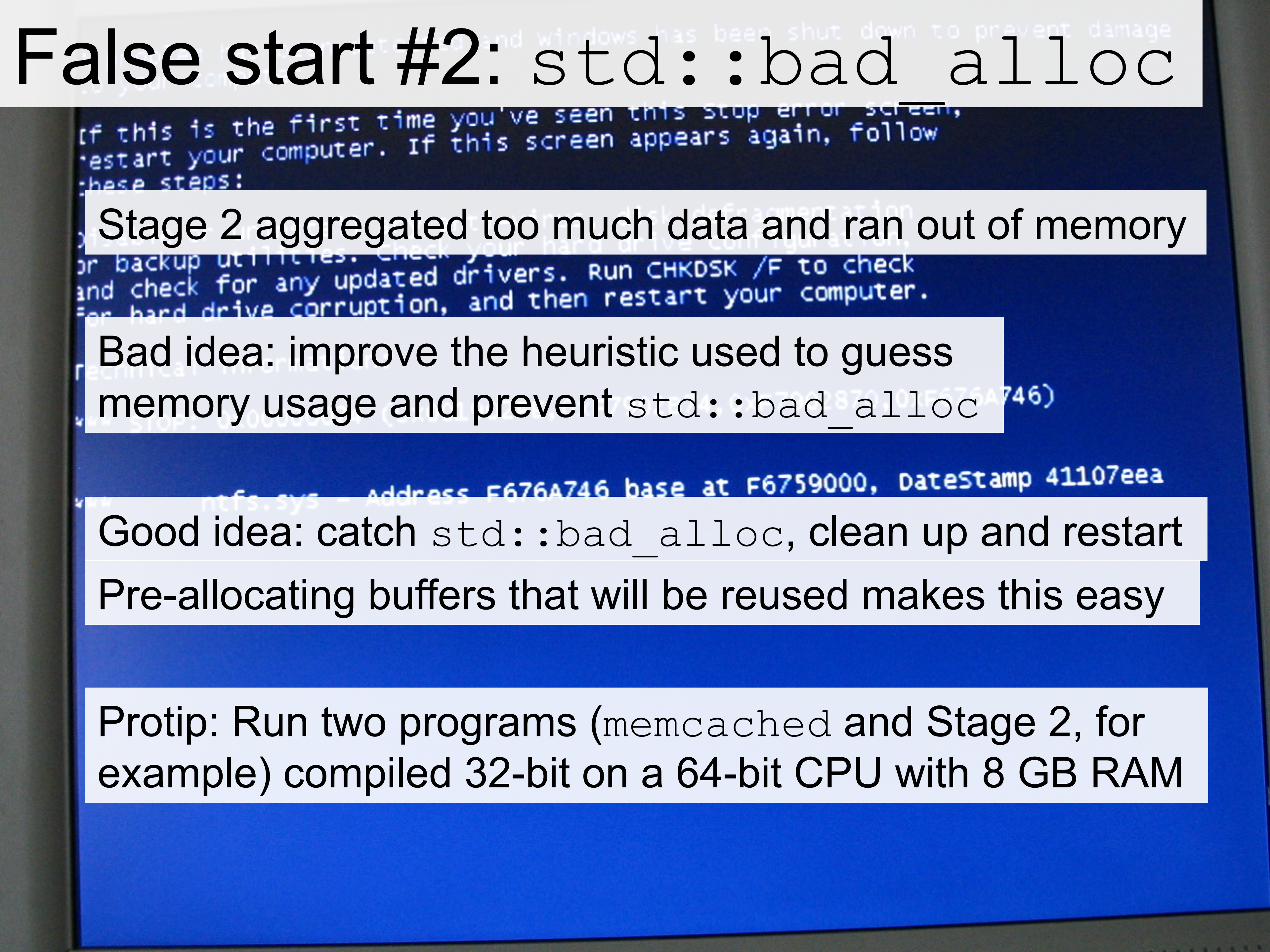
Stage 2 aggregated too much data and ran out of memory

Bad idea: improve the heuristic used to guess memory usage and prevent `std::bad_alloc`

Good idea: catch `std::bad_alloc`, clean up and restart

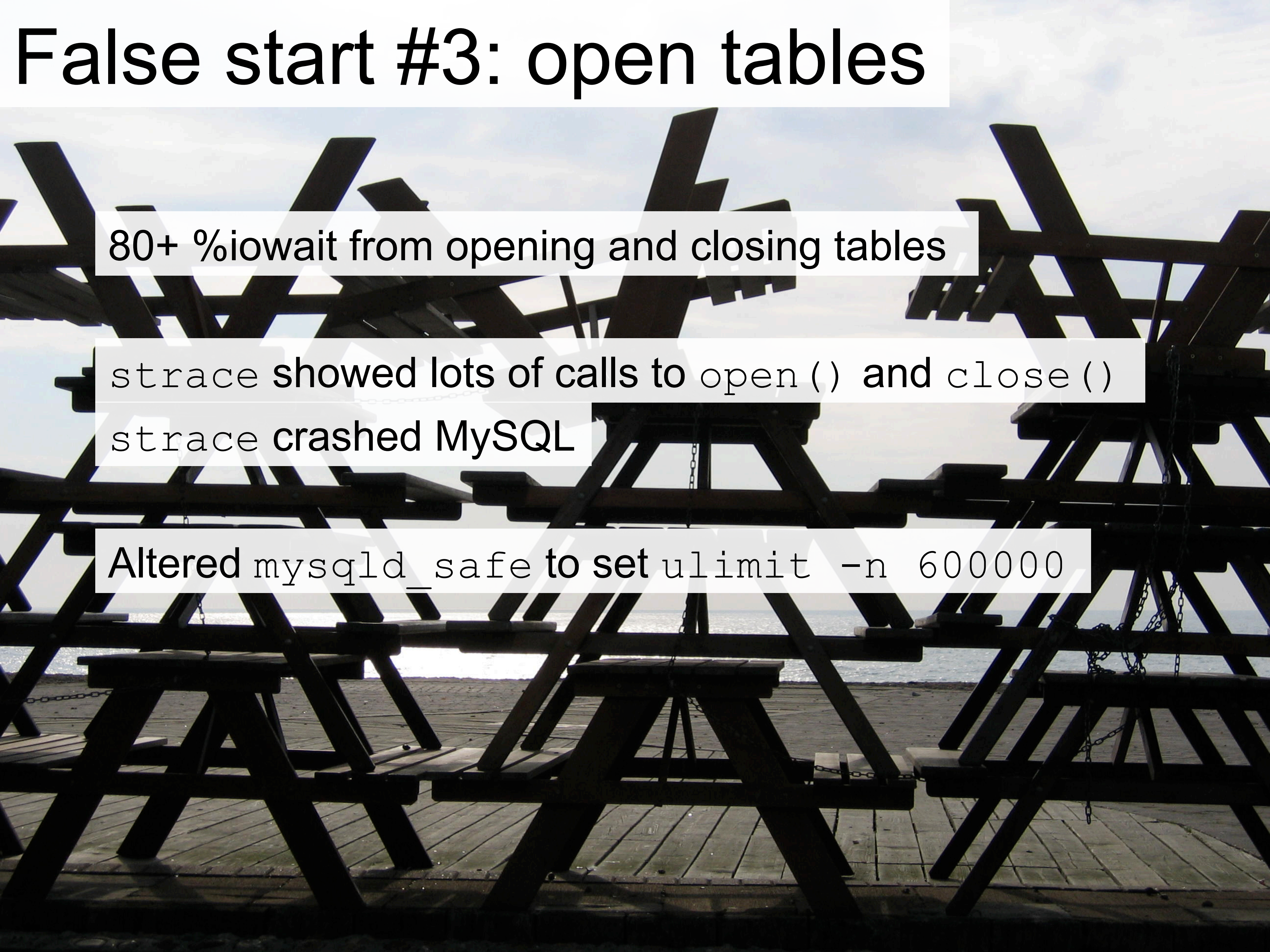
Pre-allocating buffers that will be reused makes this easy

Protip: Run two programs (`memcached` and Stage 2, for example) compiled 32-bit on a 64-bit CPU with 8 GB RAM



```
and windows has been shut down to prevent damage
f this is the first time you've seen this stop error screen,
estart your computer. If this screen appears again, follow
hese steps:

or backup utilities. Check your hard drive configuration,
and check for any updated drivers. Run CHKDSK /F to check
for hard drive corruption, and then restart your computer.

Technical information:

*** STOP: 0x00000024

*** ntfs.sys - Address F676A746 base at F6759000, DateStamp 41107eea
```

False start #3: open tables



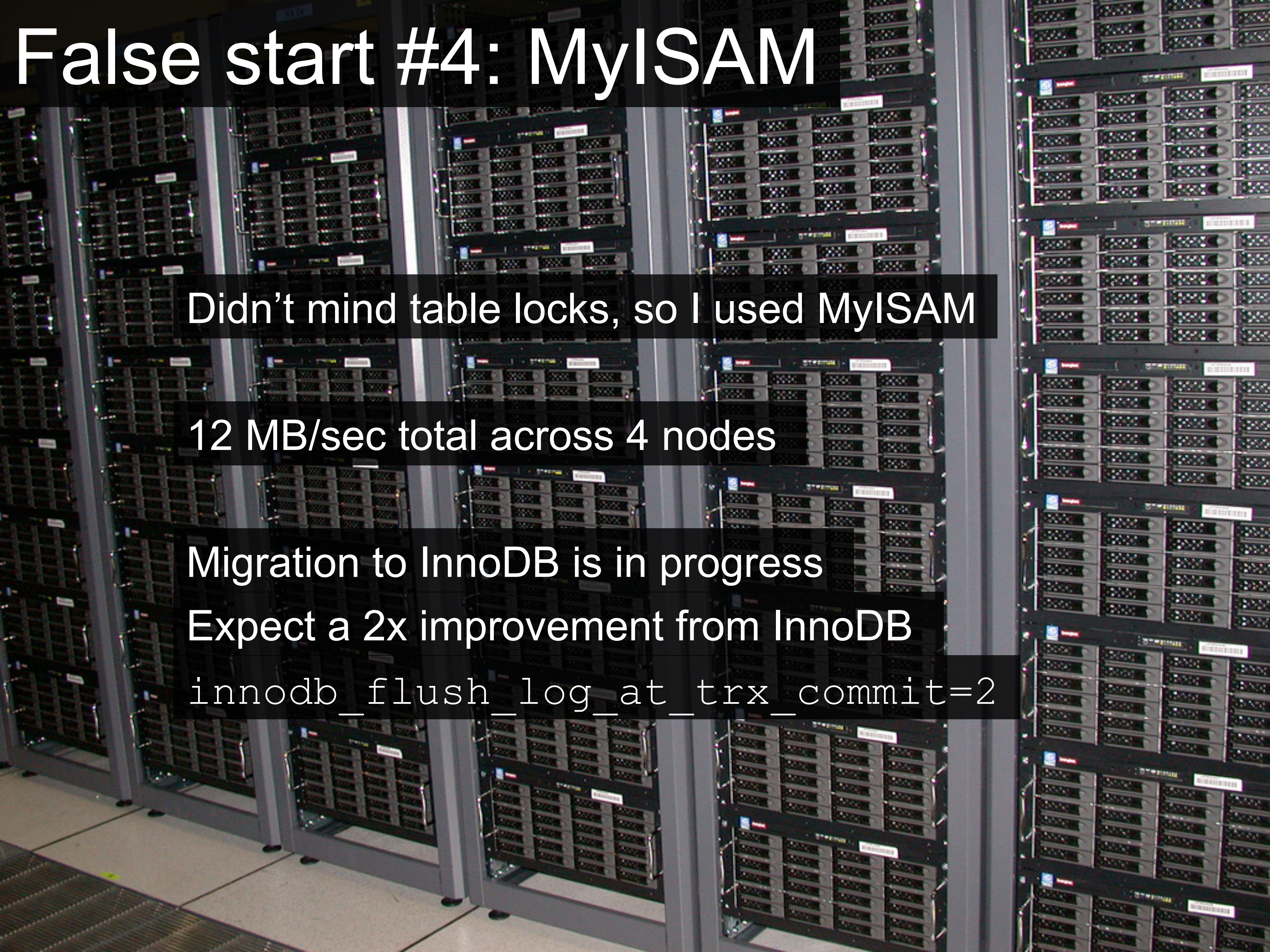
80+ %iowait from opening and closing tables

`strace` showed lots of calls to `open()` and `close()`

`strace` crashed MySQL

Altered `mysqld_safe` to set `ulimit -n 600000`

False start #4: MyISAM



Didn’t mind table locks, so I used MyISAM

12 MB/sec total across 4 nodes

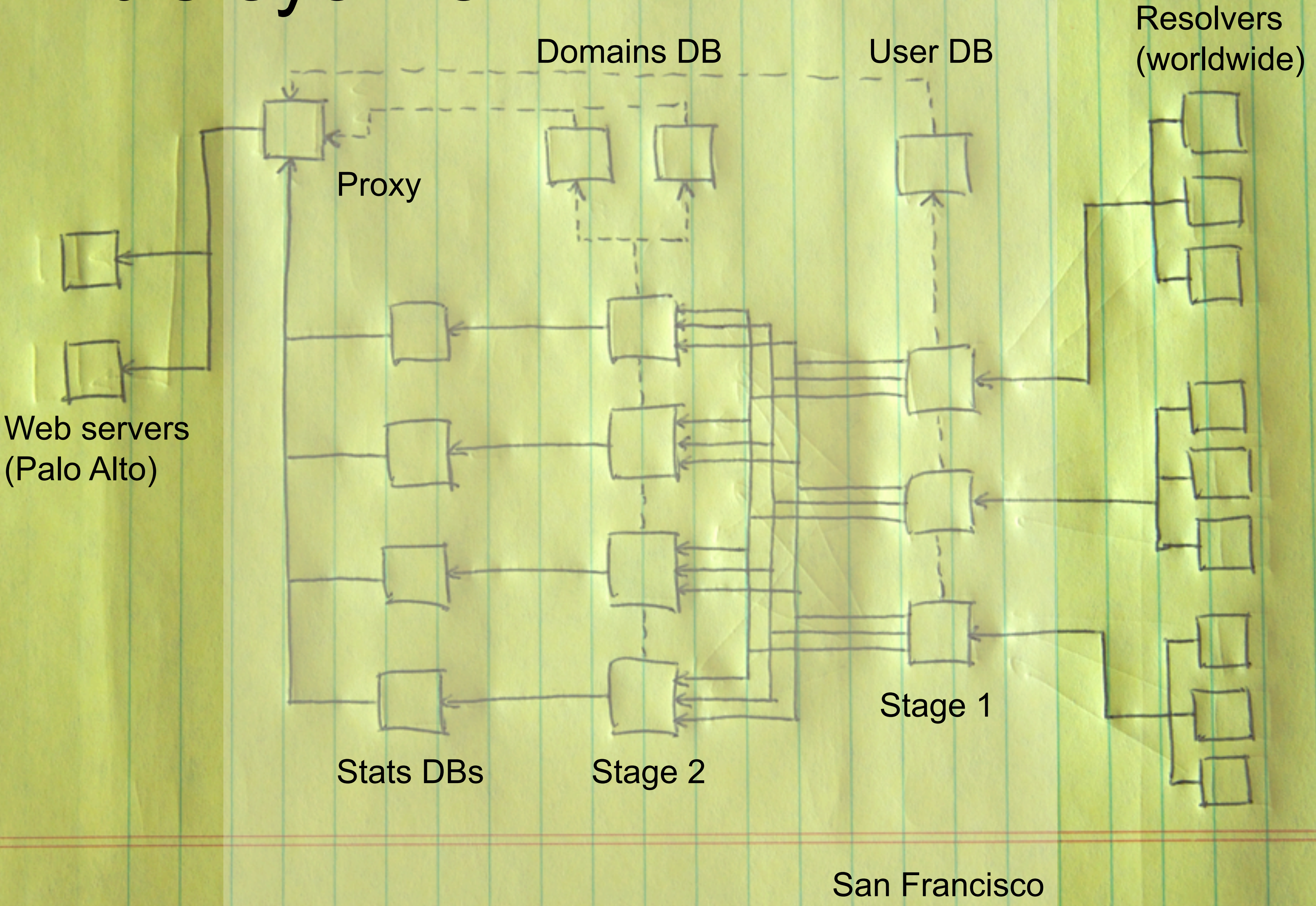
Migration to InnoDB is in progress

Expect a 2x improvement from InnoDB

```
innodb_flush_log_at_trx_commit=2
```

Architecture

Bird's eye view



Resolvers (worldwide)

Domains DB

User DB

Proxy

Web servers (Palo Alto)

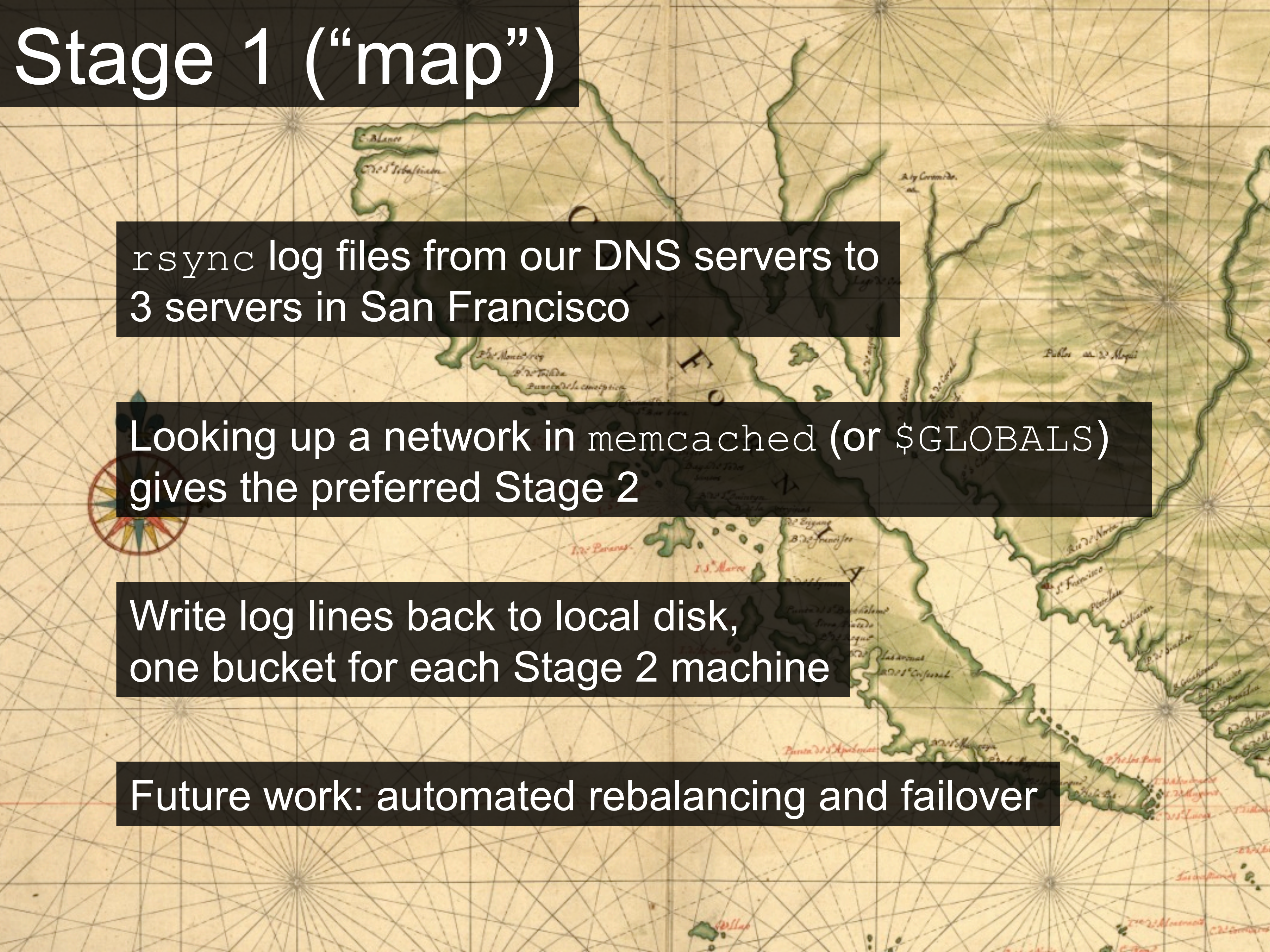
Stats DBs

Stage 2

Stage 1

San Francisco

Stage 1 (“map”)



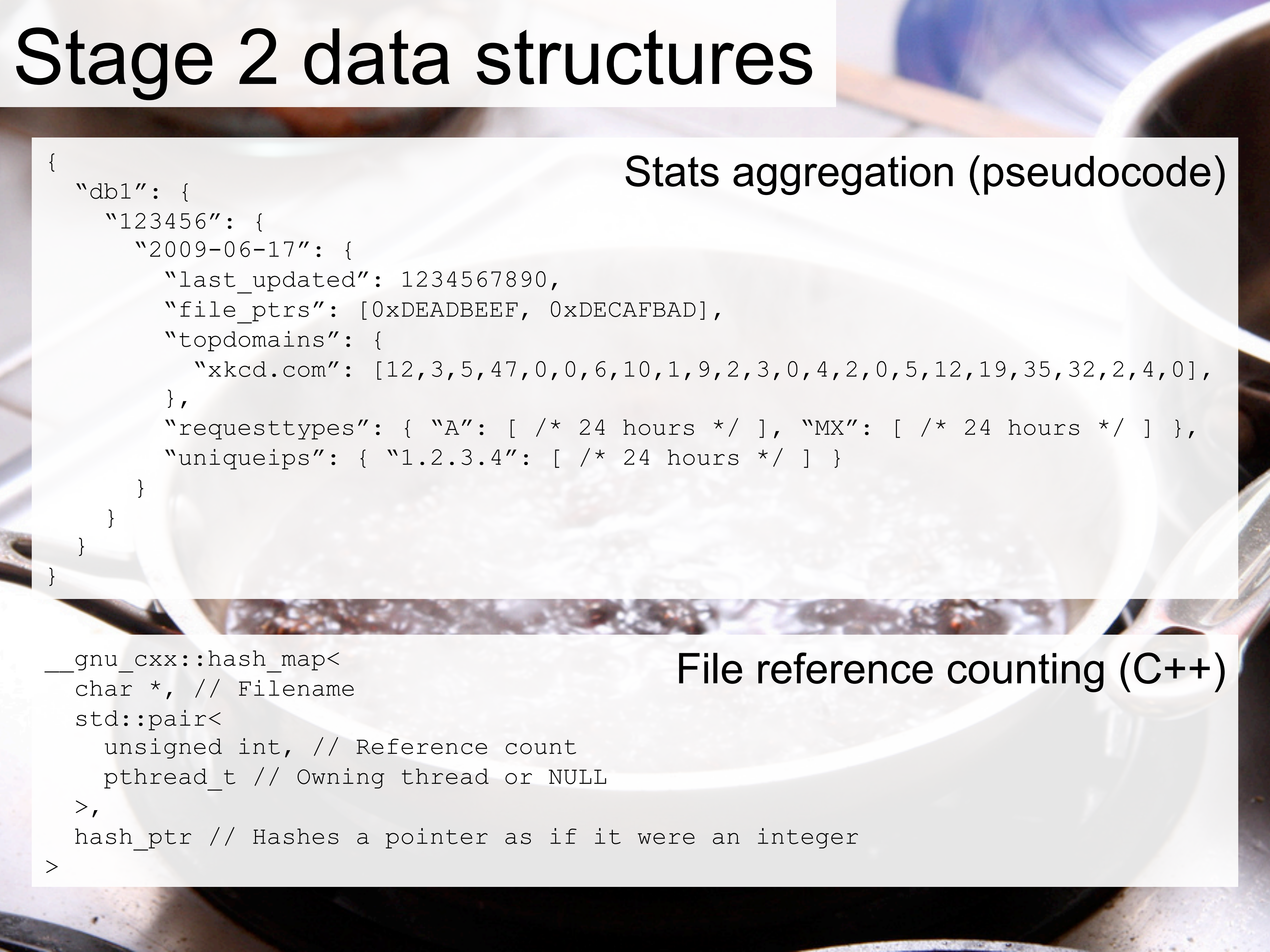
`rsync` log files from our DNS servers to 3 servers in San Francisco

Looking up a network in `memcached` (or `$GLOBALS`) gives the preferred Stage 2

Write log lines back to local disk,
one bucket for each Stage 2 machine

Future work: automated rebalancing and failover

Stage 2 data structures



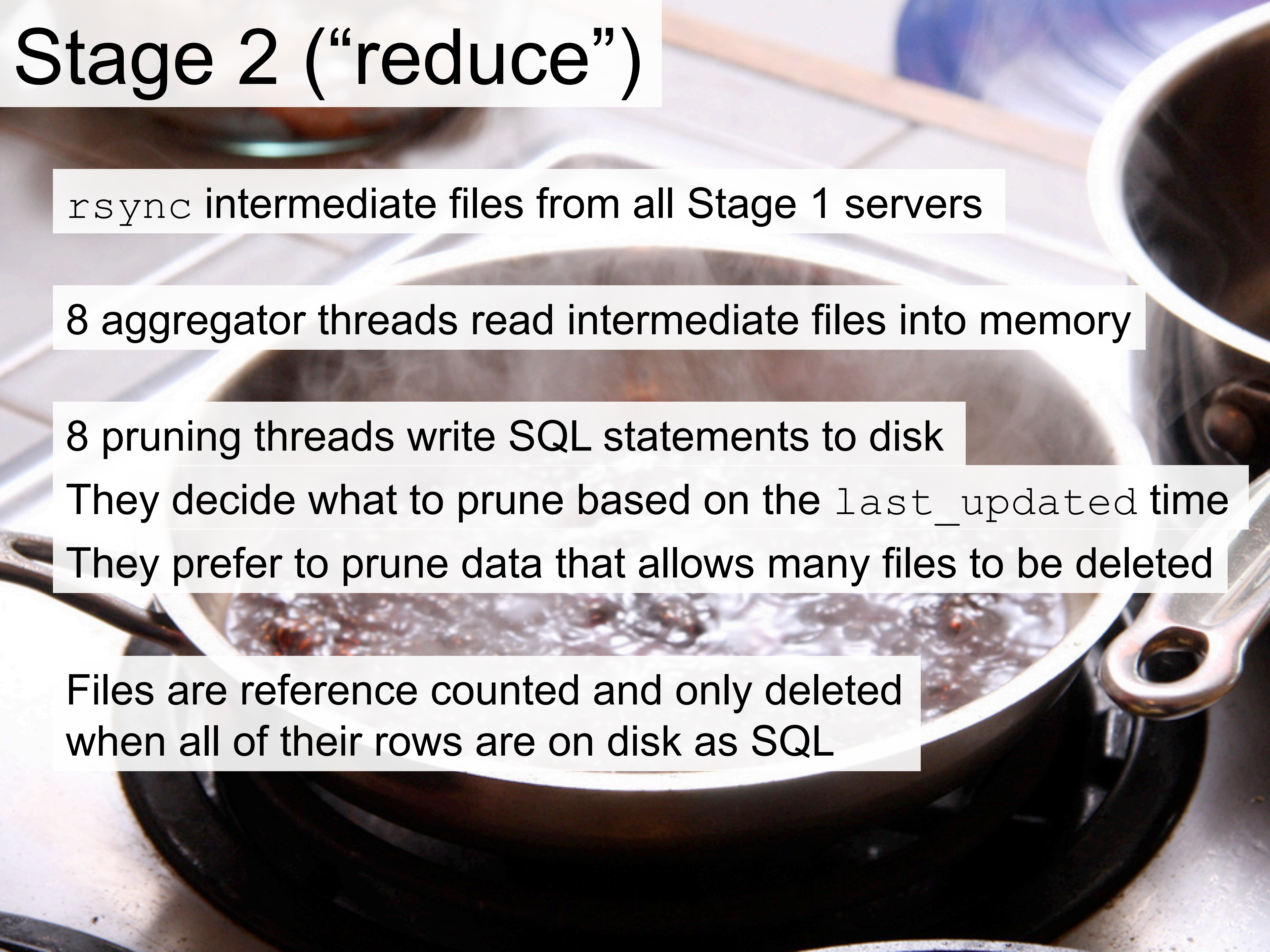
Stats aggregation (pseudocode)

```
{
  “db1”: {
    “123456”: {
      “2009-06-17”: {
        “last_updated”: 1234567890,
        “file_ptrs”: [0xDEADBEEF, 0xDECAFBAD],
        “topdomains”: {
          “xkcd.com”: [12,3,5,47,0,0,6,10,1,9,2,3,0,4,2,0,5,12,19,35,32,2,4,0],
        },
        “requesttypes”: { “A”: [ /* 24 hours */ ], “MX”: [ /* 24 hours */ ] },
        “uniqueips”: { “1.2.3.4”: [ /* 24 hours */ ] }
      }
    }
  }
}
```

File reference counting (C++)

```
__gnu_cxx::hash_map<
  char *, // Filename
  std::pair<
    unsigned int, // Reference count
    pthread_t // Owning thread or NULL
  >,
  hash_ptr // Hashes a pointer as if it were an integer
>
```

Stage 2 (“reduce”)



`rsync` intermediate files from all Stage 1 servers

8 aggregator threads read intermediate files into memory

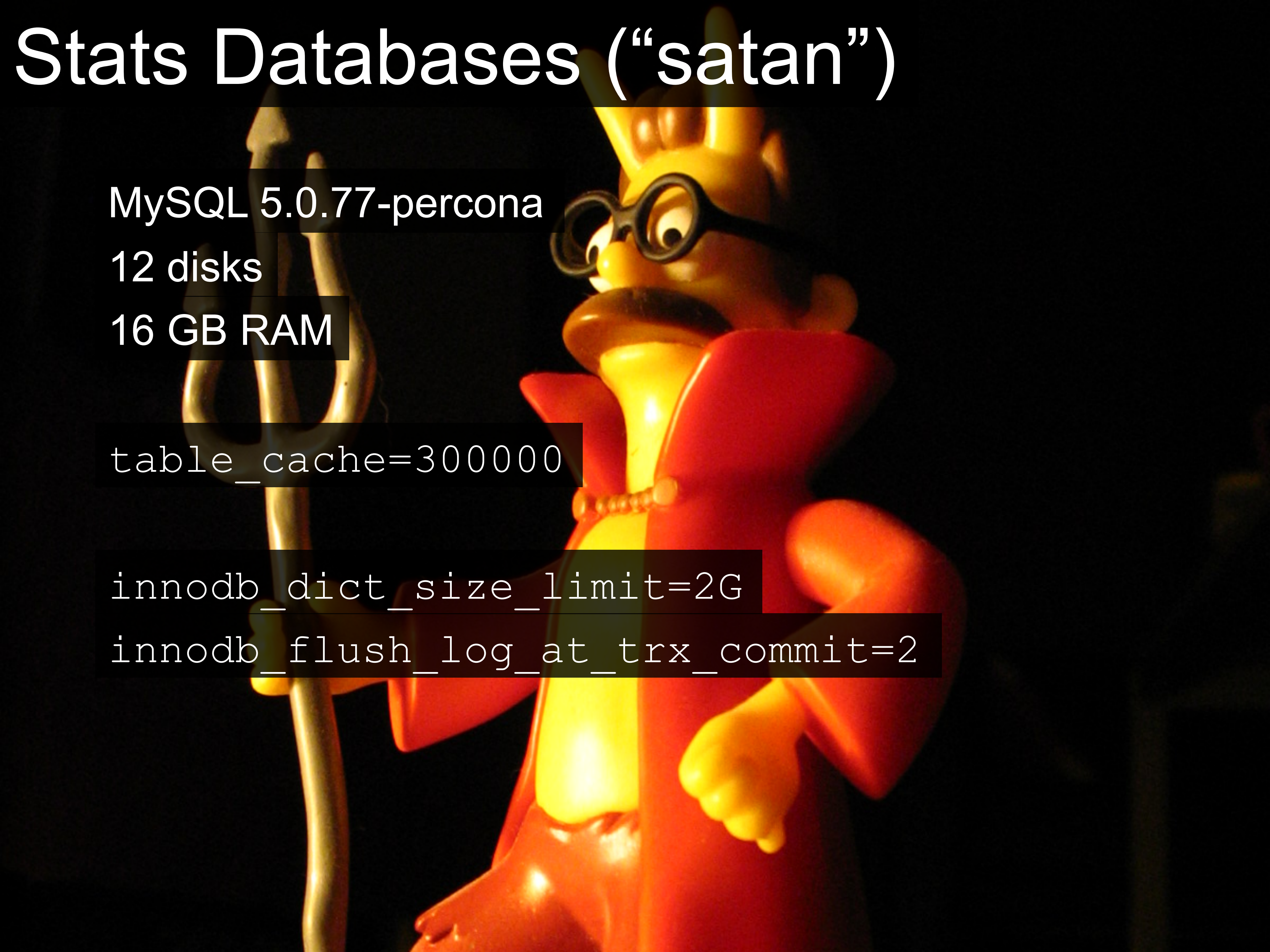
8 pruning threads write SQL statements to disk

They decide what to prune based on the `last_updated` time

They prefer to prune data that allows many files to be deleted

Files are reference counted and only deleted when all of their rows are on disk as SQL

Stats Databases (“satan”)



MySQL 5.0.77-percona

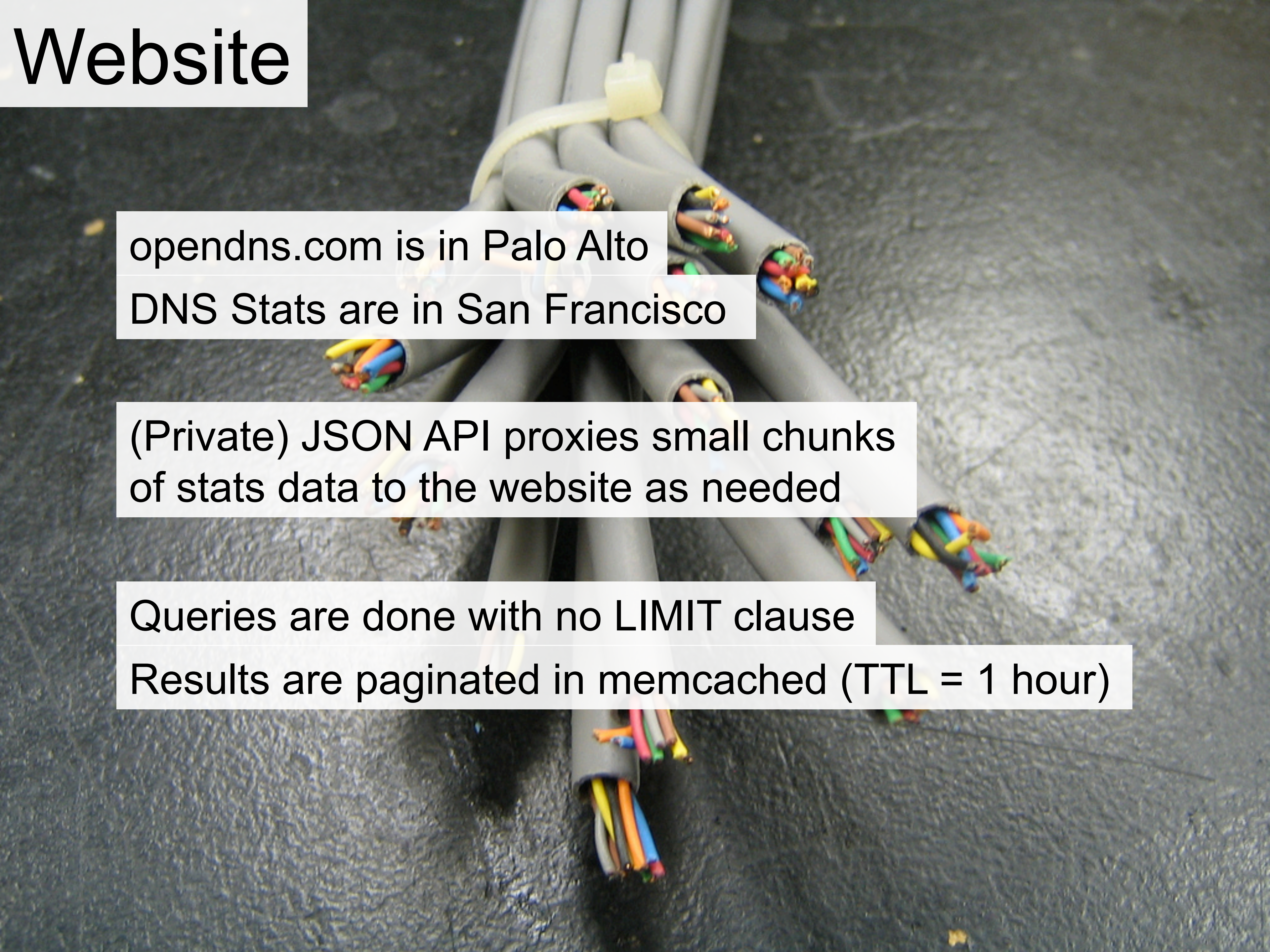
12 disks

16 GB RAM

```
table_cache=300000
```

```
innodb_dict_size_limit=2G
innodb_flush_log_at_trx_commit=2
```

Website



opendns.com is in Palo Alto

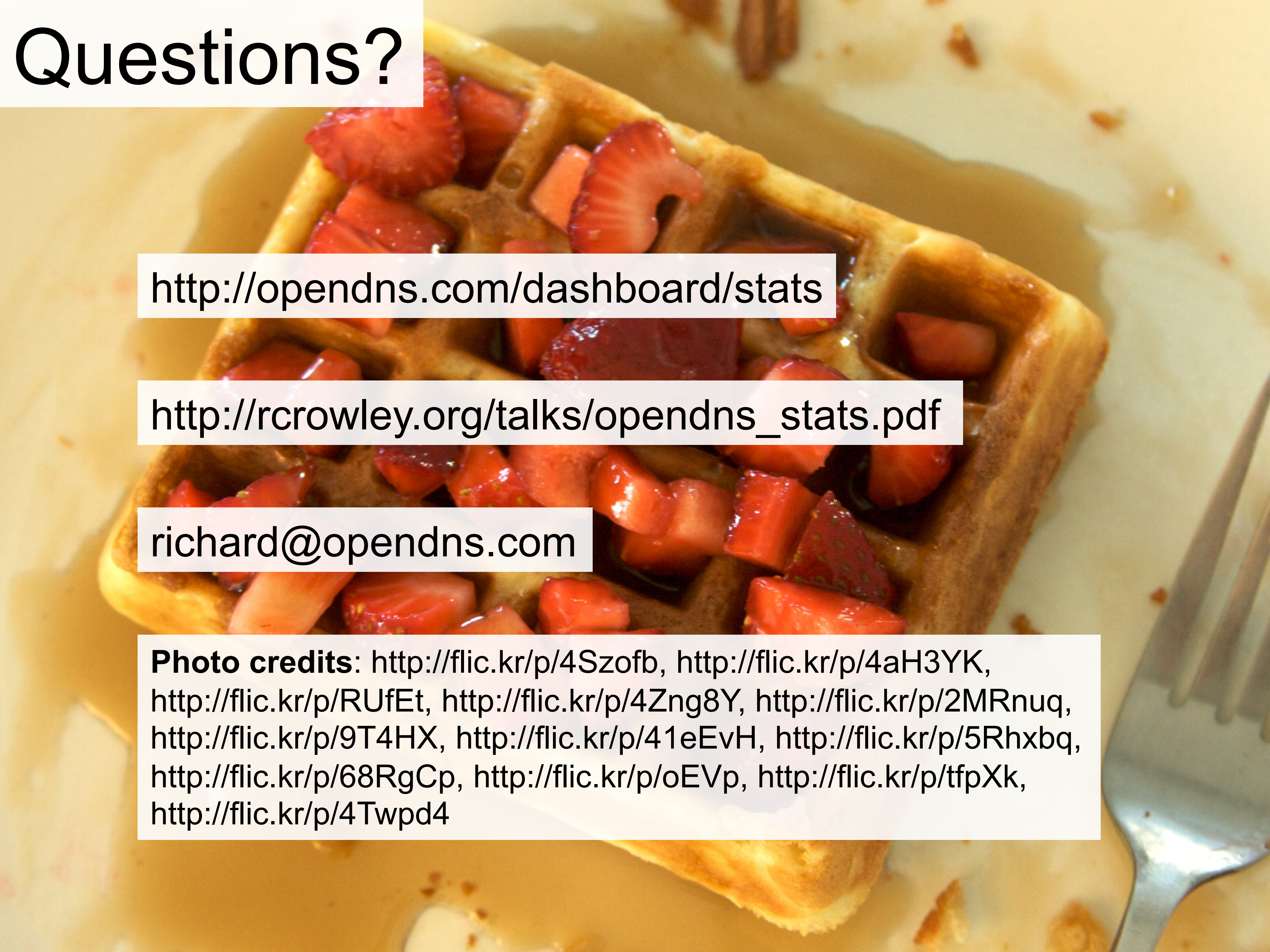
DNS Stats are in San Francisco

(Private) JSON API proxies small chunks of stats data to the website as needed

Queries are done with no LIMIT clause

Results are paginated in memcached (TTL = 1 hour)

Questions?



http://opendns.com/dashboard/stats

http://rcrowley.org/talks/opendns_stats.pdf

richard@opendns.com

Photo credits: http://flic.kr/p/4Szofb, http://flic.kr/p/4aH3YK, http://flic.kr/p/RUfEt, http://flic.kr/p/4Zng8Y, http://flic.kr/p/2MRnuq, http://flic.kr/p/9T4HX, http://flic.kr/p/41eEvH, http://flic.kr/p/5Rhxbq, http://flic.kr/p/68RgCp, http://flic.kr/p/oEVp, http://flic.kr/p/tfpXk, http://flic.kr/p/4Twpd4