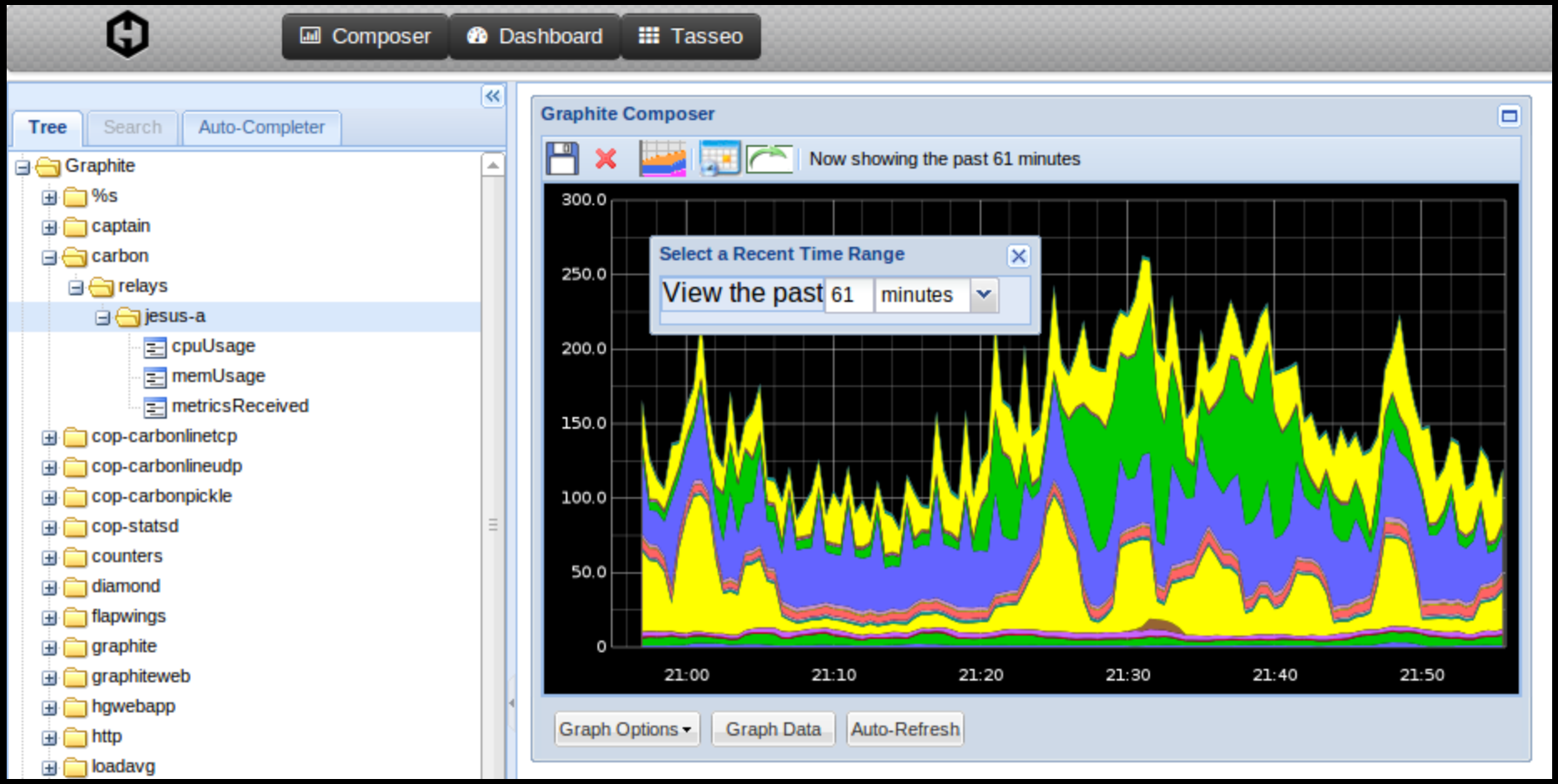
How to use Riak incorrectly

Helpful tips from Hosted Graphite

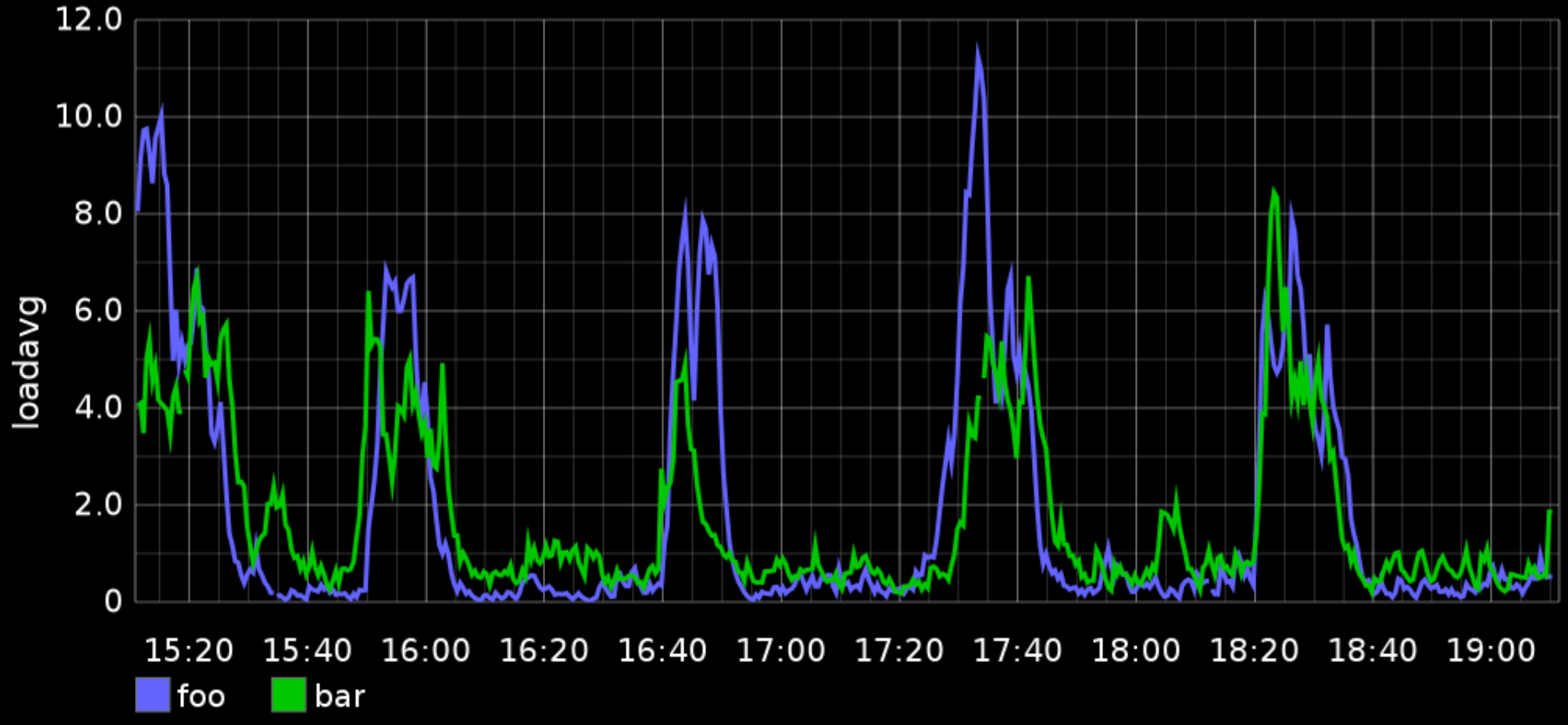
Hosted Graphite

You send us numbers, we give you graphs.


Composer
Dashboard
Tasseo
Tree
Search
Auto-Completer
Graphite
%s
captain
carbon
relays
jesus-a
cpuUsage
memUsage
metricsReceived
cop-carbonlinetcp
cop-carbonlineudp
cop-carbonpickle
cop-statsd
counters
diamond
flapwings
graphite
graphiteweb
hgwebapp
http
loadavg
Graphite Composer
Now showing the past 61 minutes
Select a Recent Time Range
View the past 61 minutes
300.0
250.0
200.0
150.0
100.0
50.0
0
21:00
21:10
21:20
21:30
21:40
21:50
Graph Options
Graph Data
Auto-Refresh

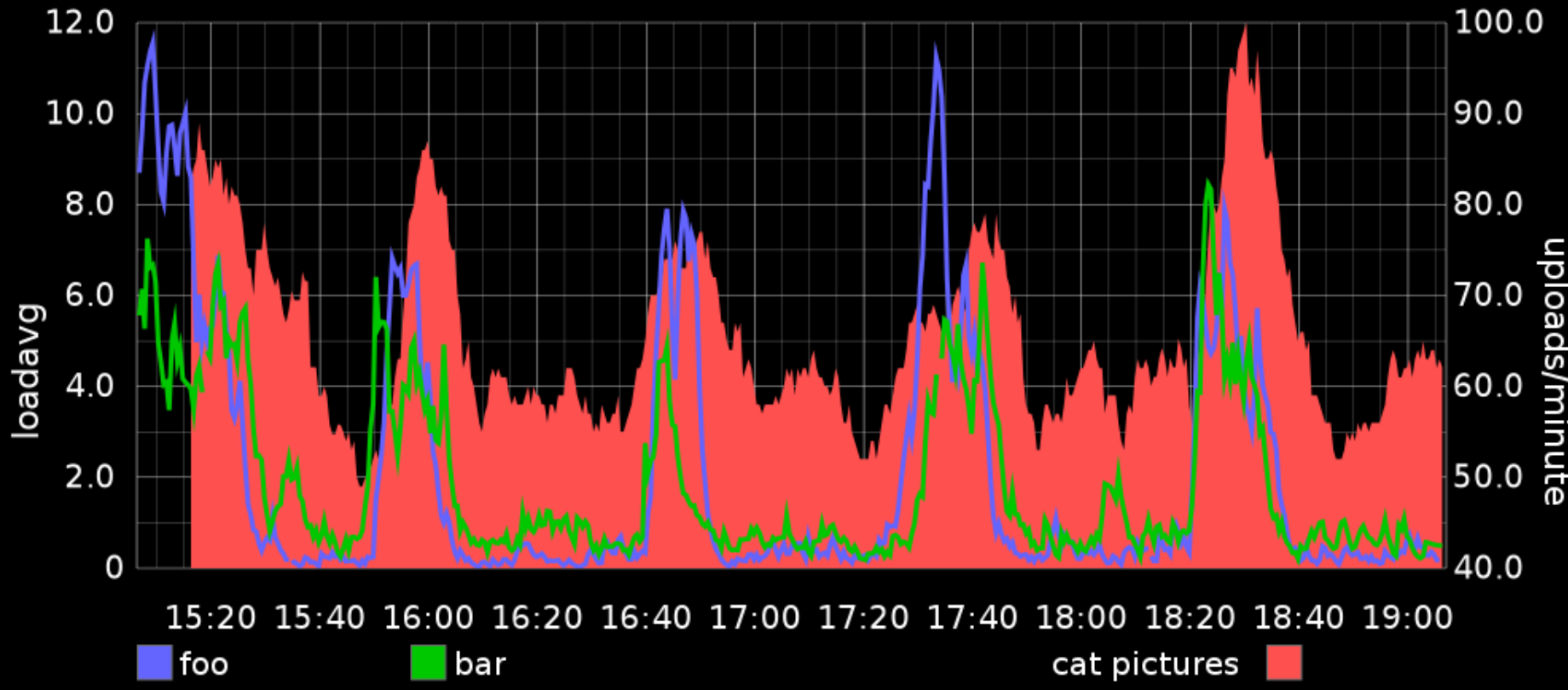
"servers.foo.loadavg 0.34"
"servers.bar.loadavg 1.40"

"servers.foo.loadavg 0.34"
"servers.bar.loadavg 1.40"

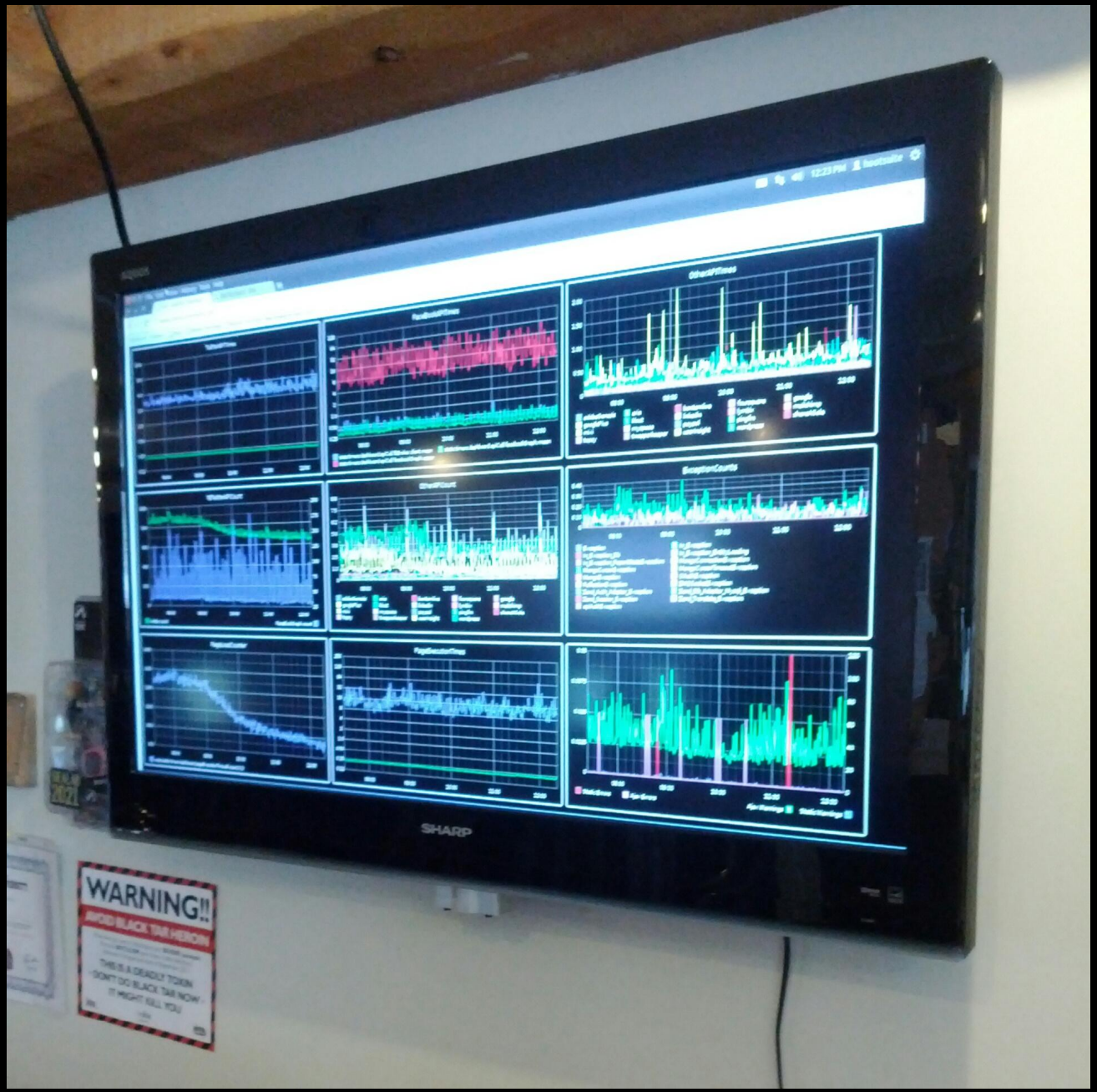


"servers.foo.loadavg 0.34"
"servers.bar.loadavg 1.40"
"catpictures.uploaded 61"

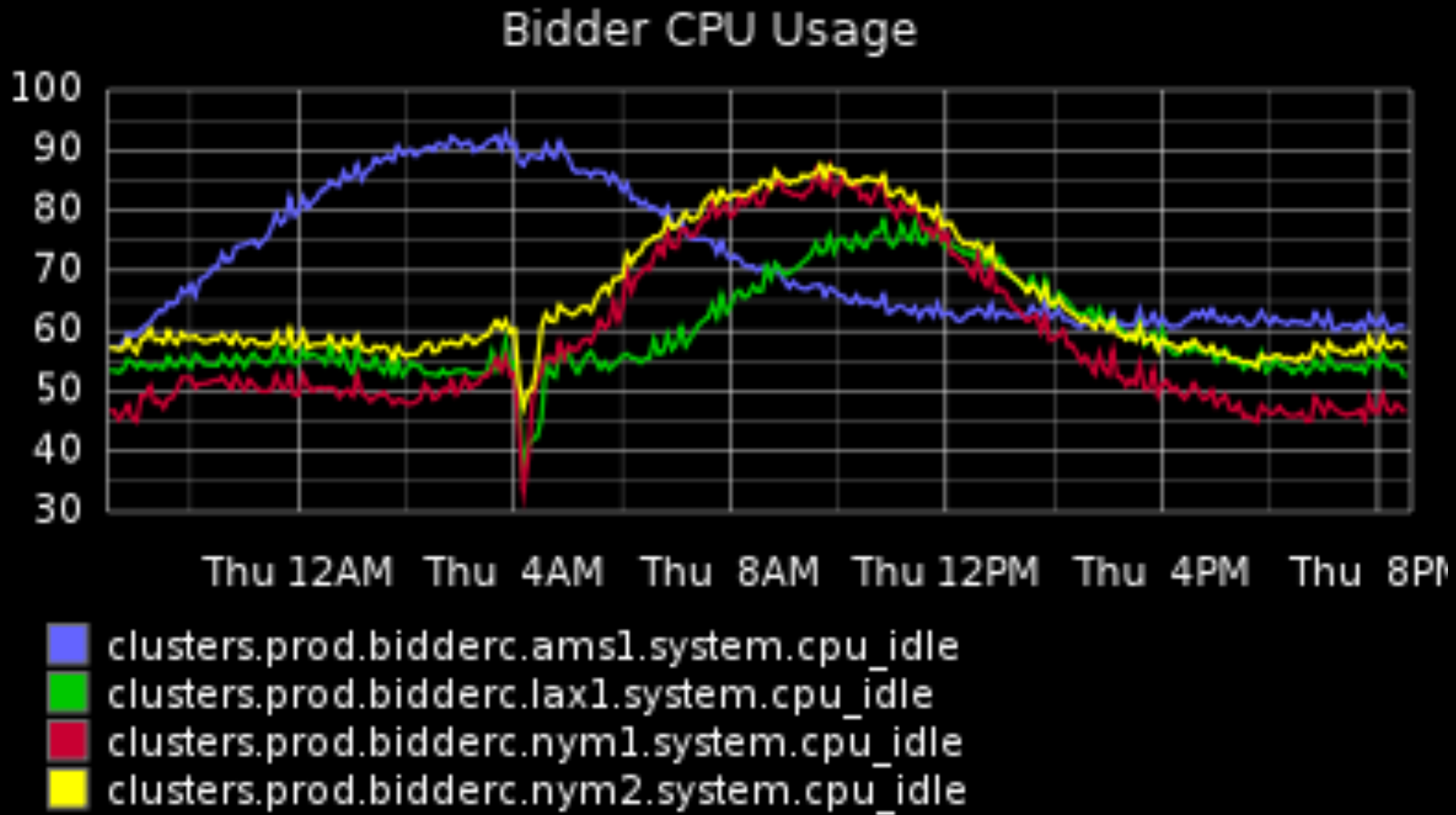
"servers.foo.loadavg 0.34"
"servers.bar.loadavg 1.40"
"catpictures.uploaded 61"



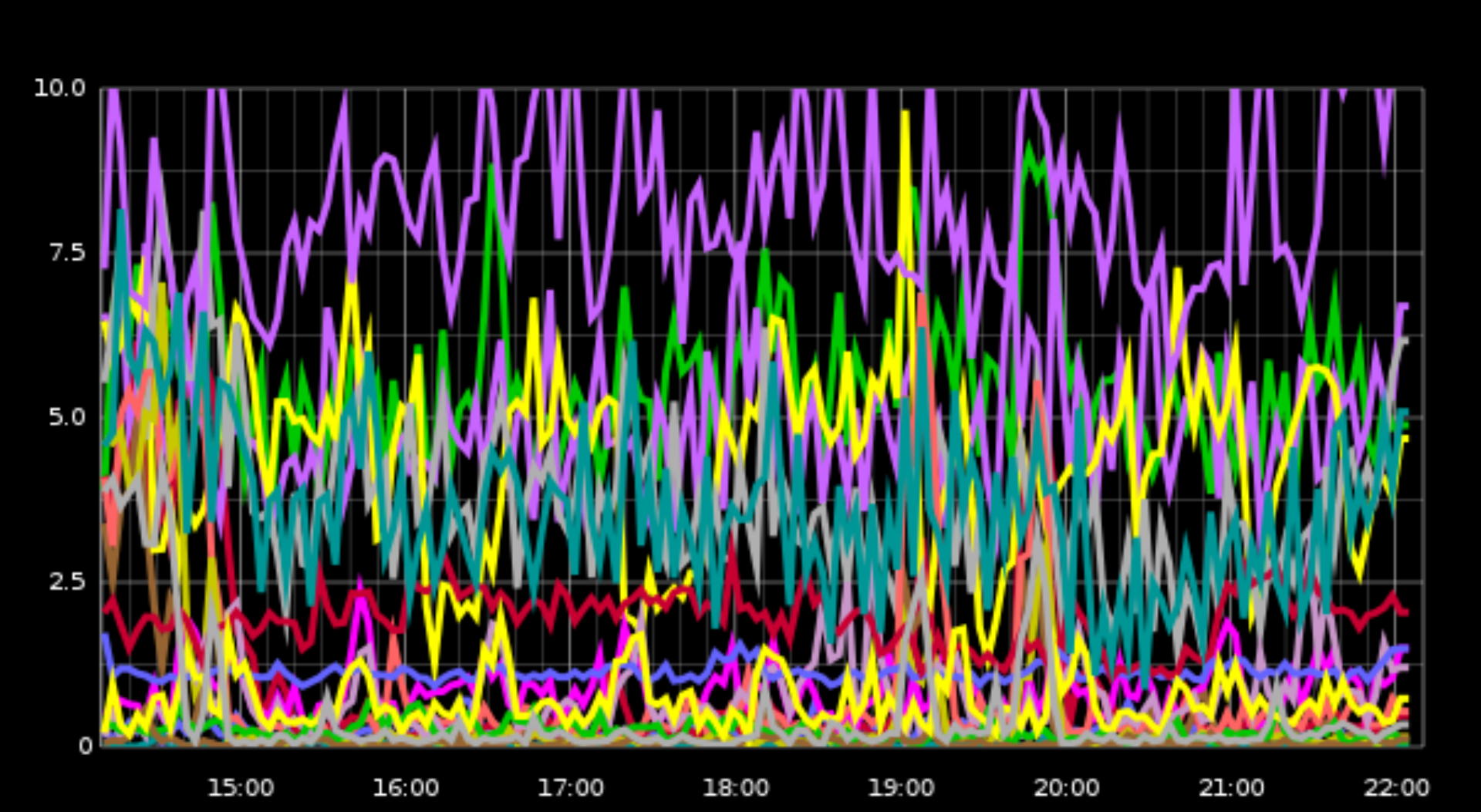
Source: changelog.ca (HootSuite)



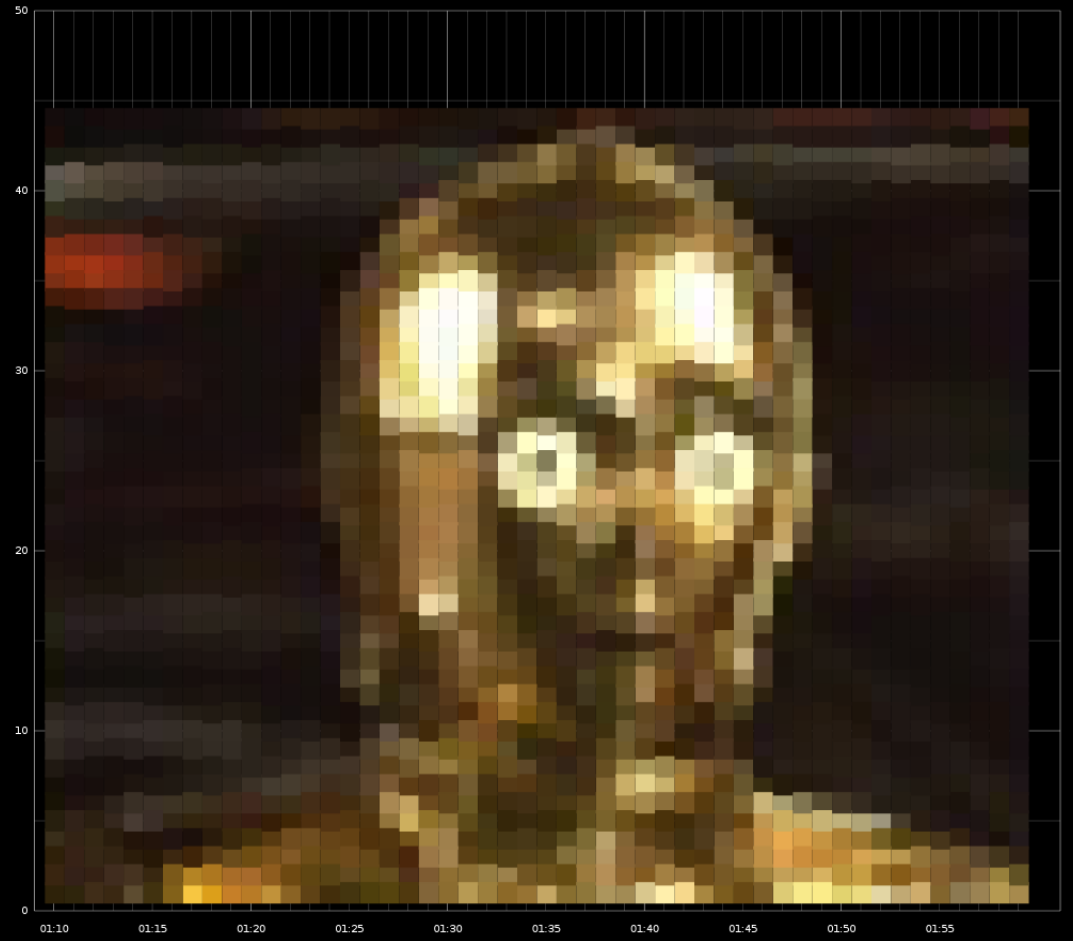
Source: changelog.ca (HootSuite)



Source: techblog.appnexus.com (AppNexus)


10.0
7.5
5.0
2.5
0
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00

Source: blog.matthewskelton.net



Source: Me!

Hosted Graphite

You send us numbers, we give you graphs.

Graphite as a Service

Large time series database as a service
... with an open source front end

authentication, account sharing, statsd, etc

Graphite as a Service

Photo by Tom Mortimer April 2011
http://www.mindatnh.org/Graphite%20Gallery1.html
http://www.mindatnh.org/gallery%20photos/Graphite%2015.jpg

1.6 million metrics stored

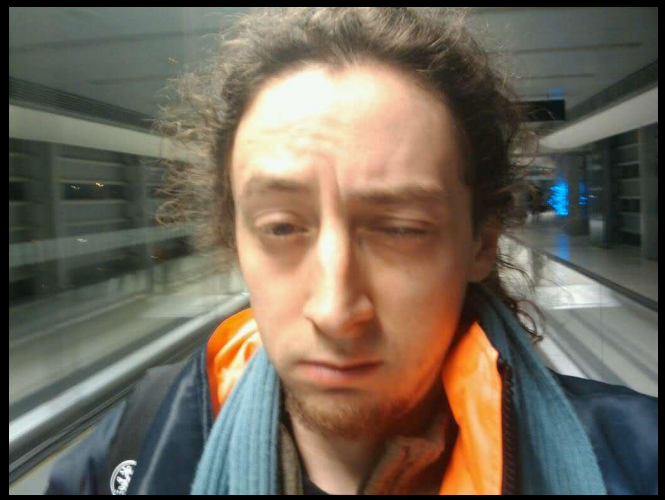
servers.foo.loadavg

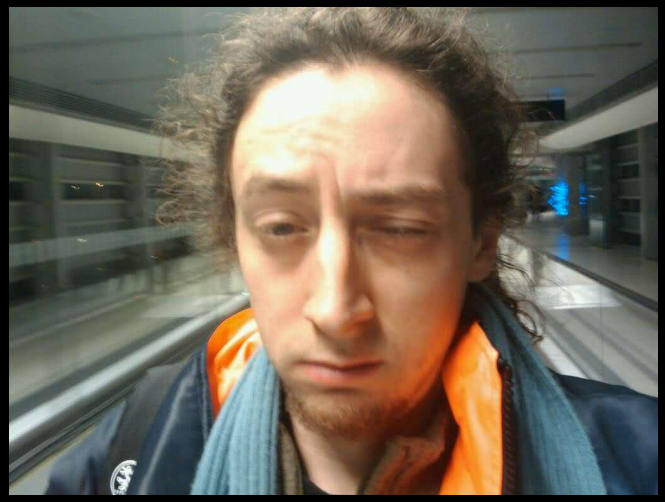
>1.5 billion datapoints/day

"servers.foo.loadavg 1.40"

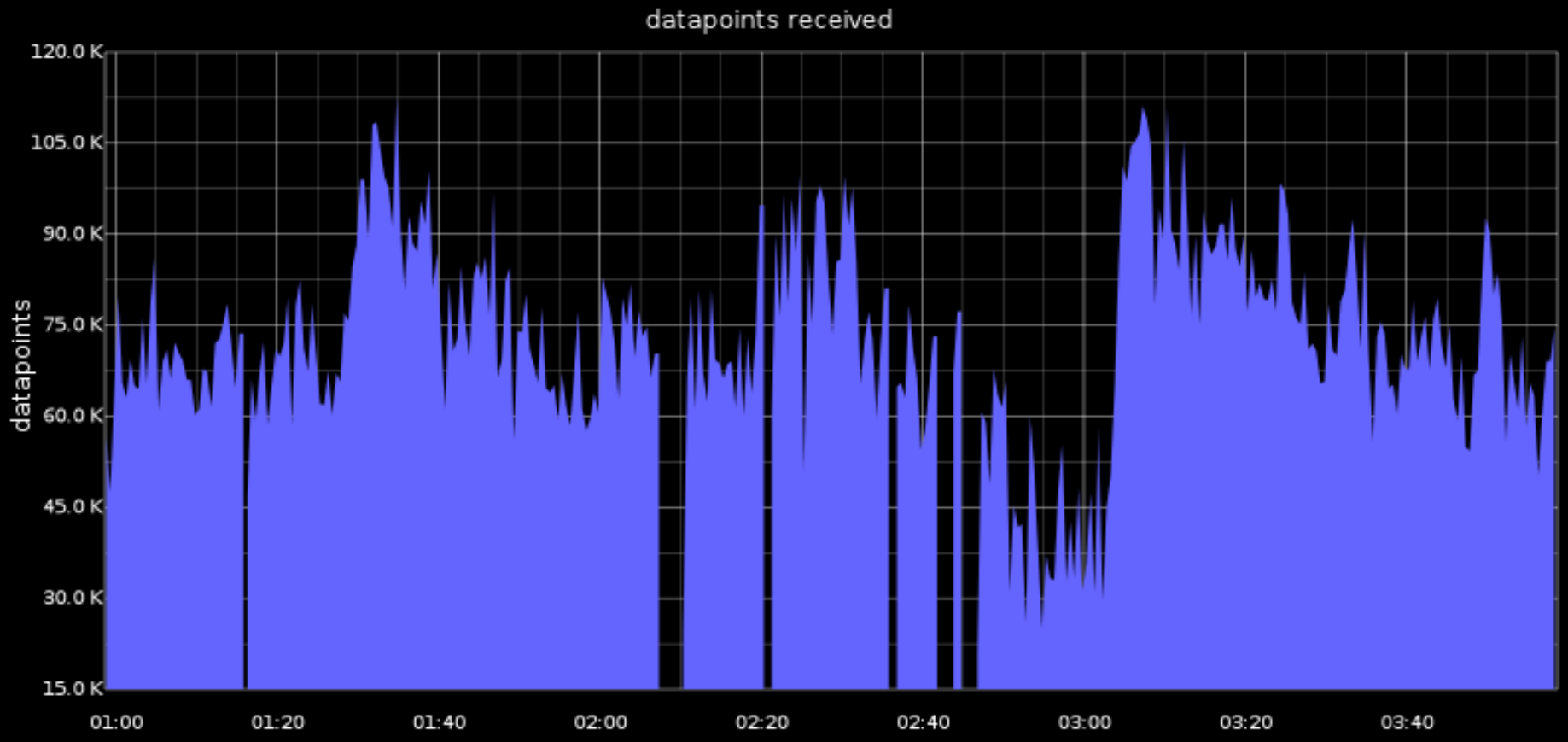
riak?

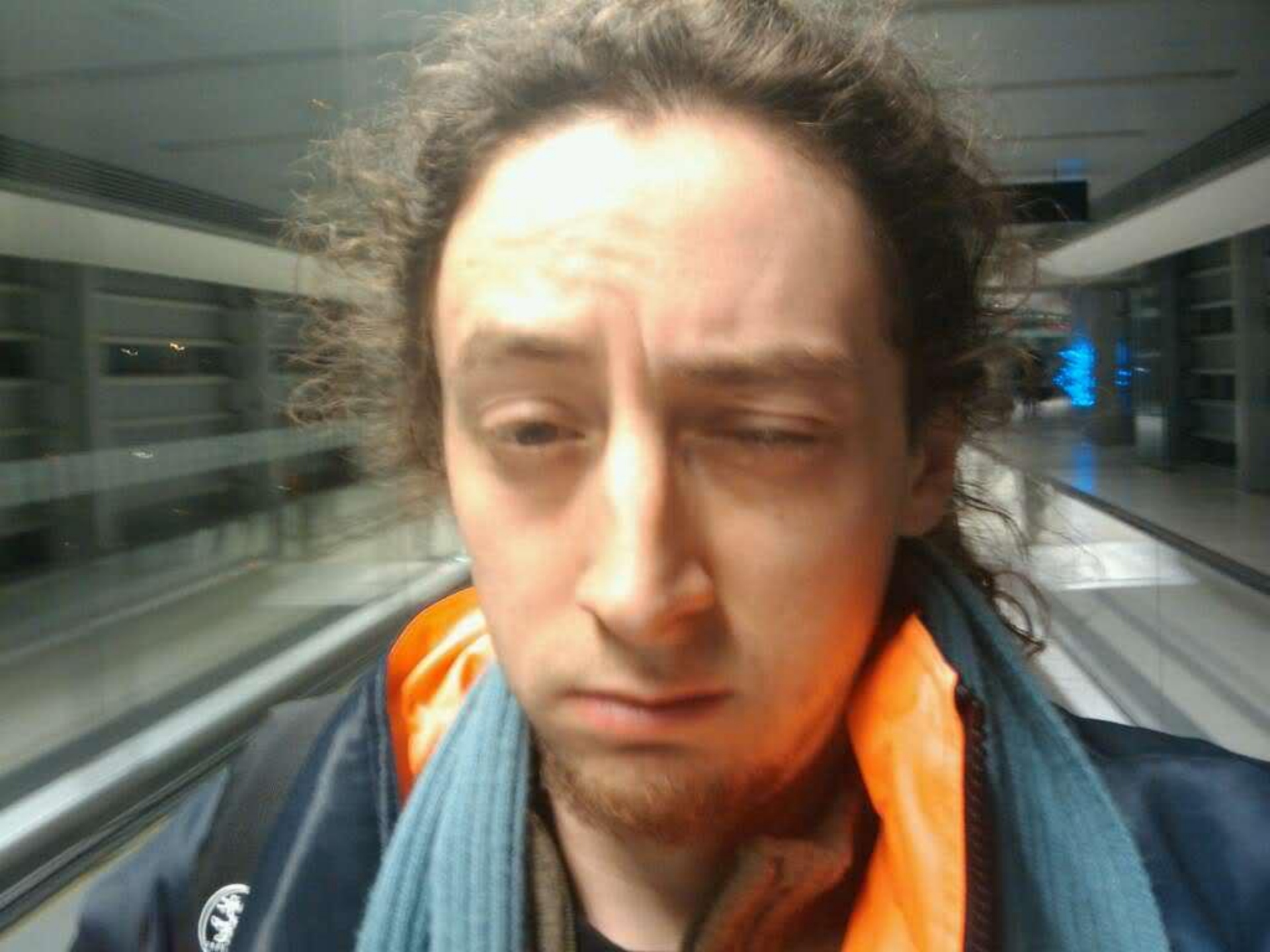
riak?





Founders: 2
Investment: €0, $0, ฿0
Failures: highly visible and can't hide them later

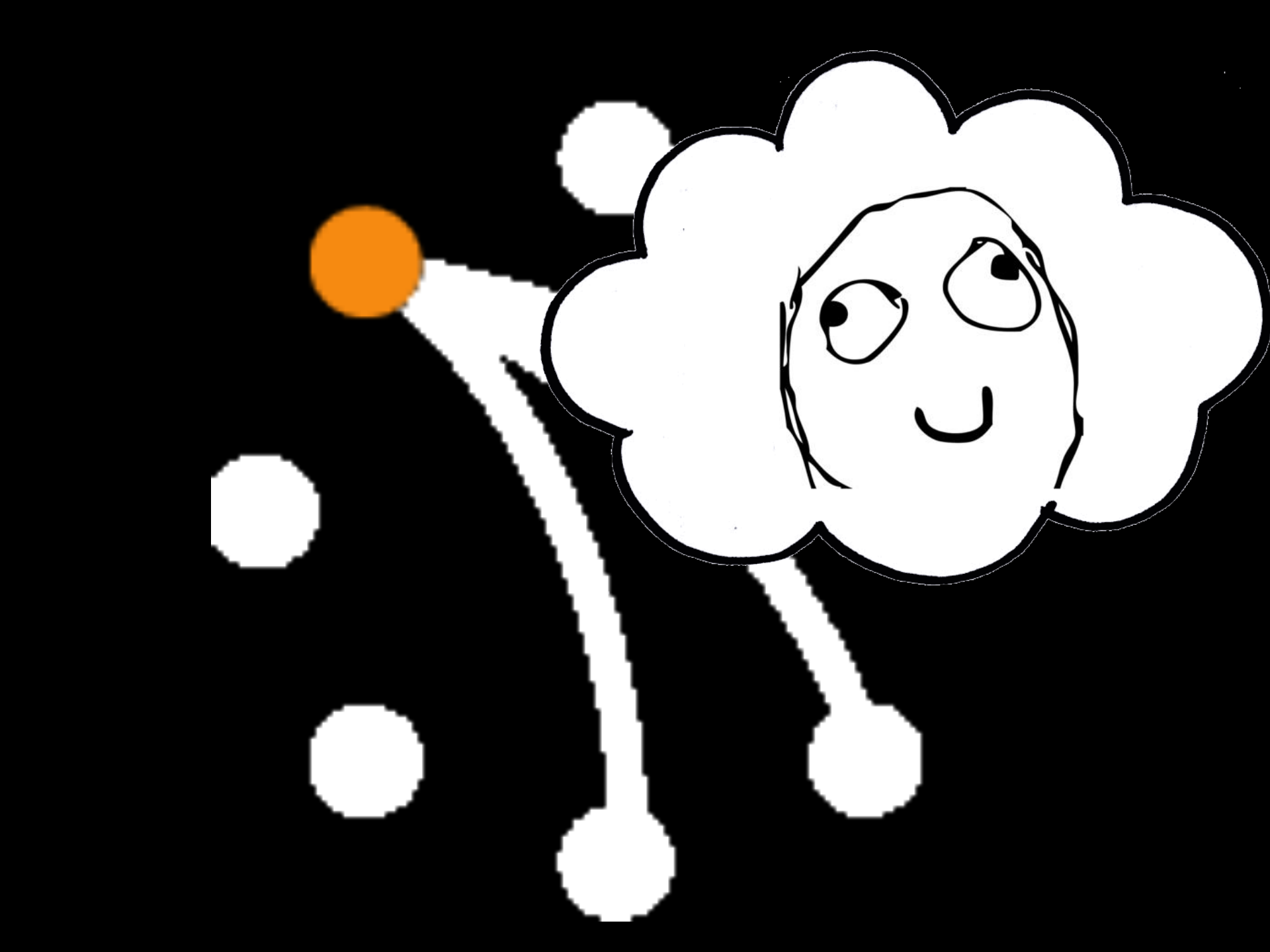

datapoints received
120.0 K
105.0 K
90.0 K
75.0 K
60.0 K
45.0 K
30.0 K
15.0 K
datapoints
01:00
01:20
01:40
02:00
02:20
02:40
03:00
03:20
03:40

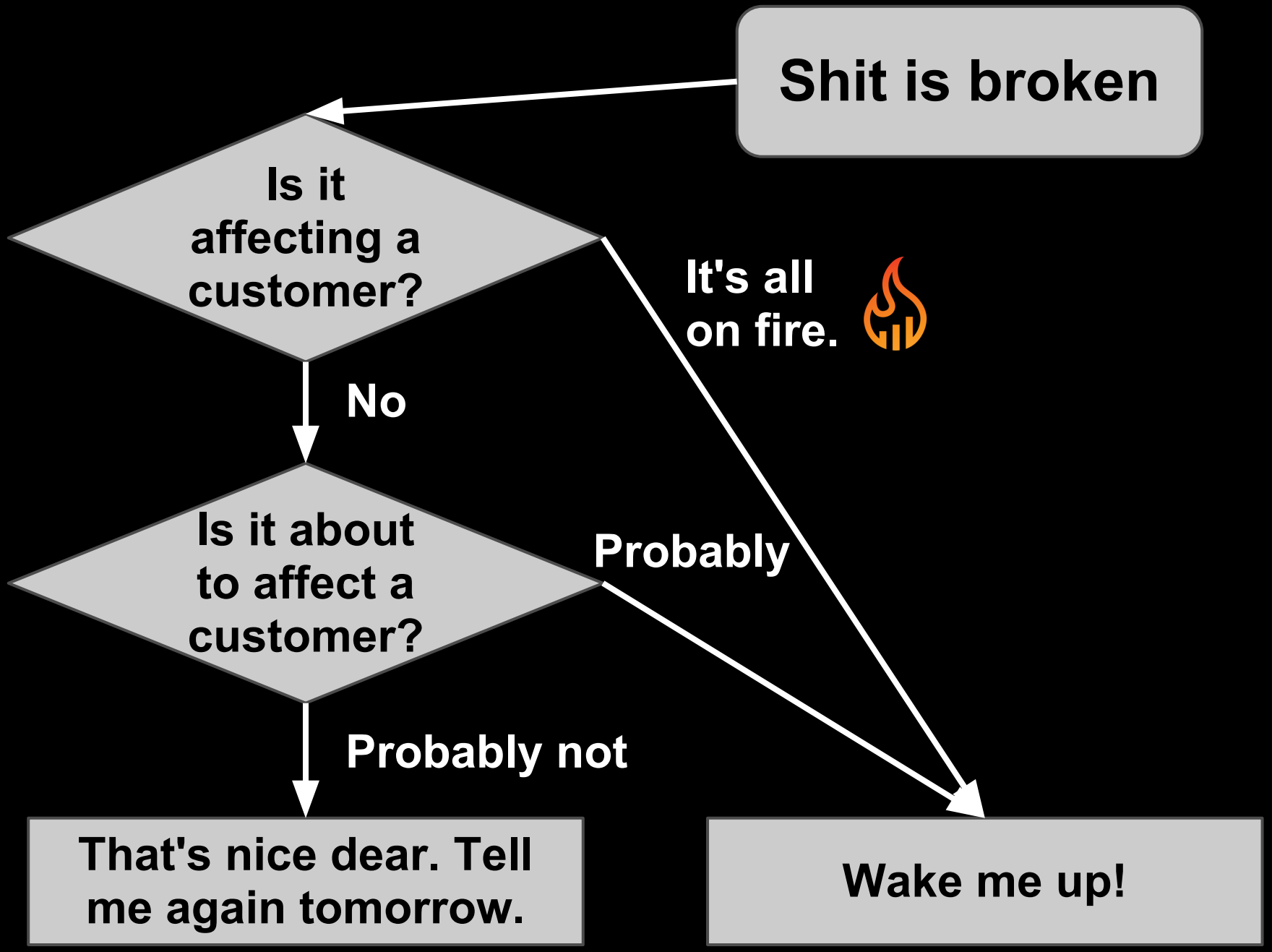


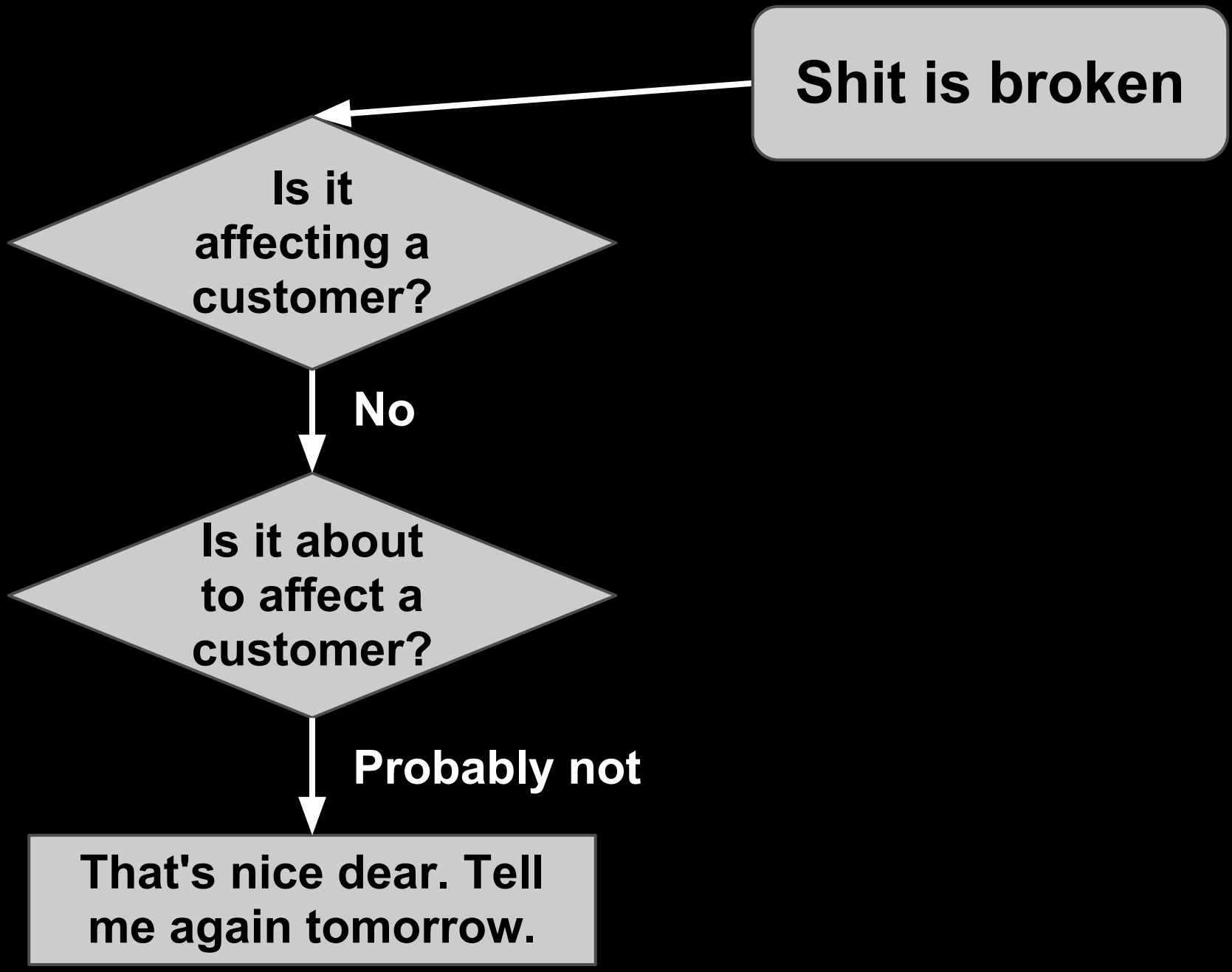
riak

helps us

sleep




Shit is broken
Is it affecting a customer?
It's all on fire.
No
Is it about to affect a customer?
Probably
Probably not
That's nice dear. Tell me again tomorrow.
Wake me up!


Shit is broken
Is it affecting a customer?
No
Is it about to affect a customer?
Probably not
That's nice dear. Tell me again tomorrow.

Shit is broken

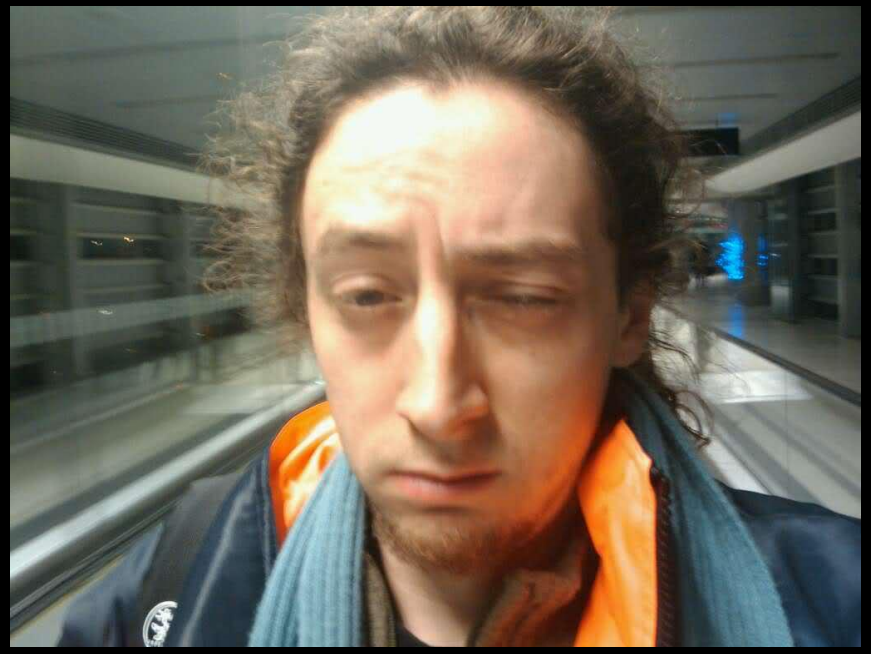
Is it affecting a customer?

No

Is it about to affect a customer?

Probably not

That's nice dear. Tell me again tomorrow.



Failures where Riak (mostly) doesn't screw you

Failing disk

High network latency

High/complete network packet loss

Out of memory

Memory corruption

Lost 2 out of 4 nodes

Unpleasant latency, but service stayed up!

riak

helps us

sleep

Things we like about Riak

... and some of the tradeoffs

Adding and removing nodes

No reconfiguring sharding, migrating data, chewing fingernails, babysitting, etc

Adding a node

```
# apt-get install riak

# /etc/init.d/riak stop
# vim /etc/riak/{vm.args,app.config}
# /etc/init.d/riak start

# riak-admin join riak@hostN.example.com
```

No control of data location

Working with cluster-level units.
(or backend-level units...)

Eventual Consistency

Split-brain? Meh, just wait.

Siblings

Merge/resolution code everywhere

No node is special

No master/slave configurations!
Shoot any node in the head!
Shoot **many** nodes!
It's all good!

All nodes are equal

All nodes *have to be* equal

Same spec, and lots of them. Expensive.

Multi-datacentre replication

Part of our disaster recovery plan

Not FOSS.

Enterprise features, enterprise pricing!

How to use Riak incorrectly

Helpful tips from Hosted Graphite

Deleting stuff

An epic adventure

Random access on spinning rust sucks

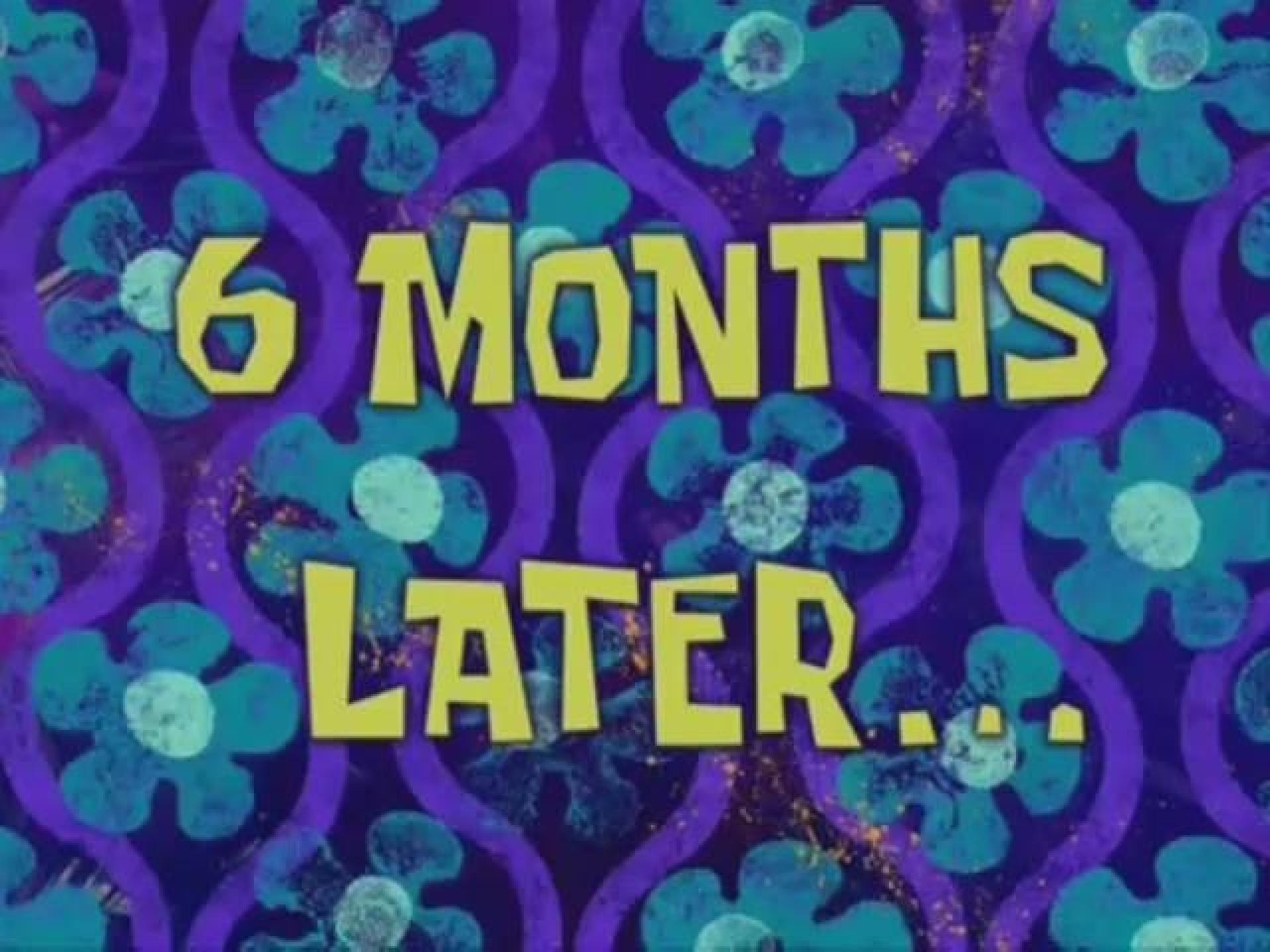
Let's just append, that's fast!

Deleting is O(expensive)

Regularly rewriting huge files to garbage collect/compact

We didn't delete.

Some lean startup nonsense, not even sure we have a product yet...


6 MONTHS
LATER...

The product works!

Consuming 15gb/day/node.

Good thing we have a delete plan!

Glad we used secondary indexes!

Otherwise...

Key listing

Just because you can...

Never do this.

10m keys?

50 bytes per key = 476.84mb

476.84mb

* (n=3)

476.84mb

* (n=3) = 1,428mb internally

476.84mb

* (n=3) = 1,428mb internally
without serialisation overhead
... or any overhead

Not for production use

This operation requires traversing all keys stored in the cluster and should not be used in production.

Glad we used 2i!

Key listings are bad.

Plan A

Delete some old stuff

2i lookup, delete, delete, ...

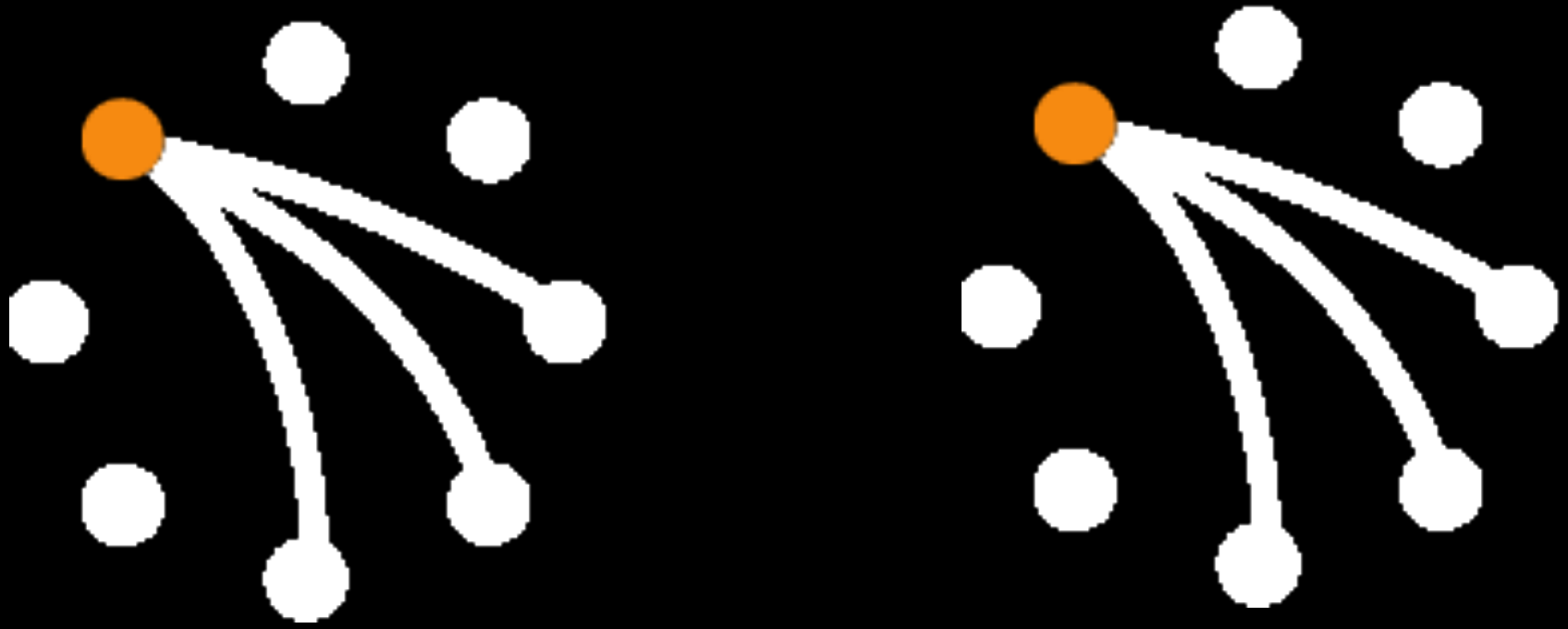
Can't keep up.

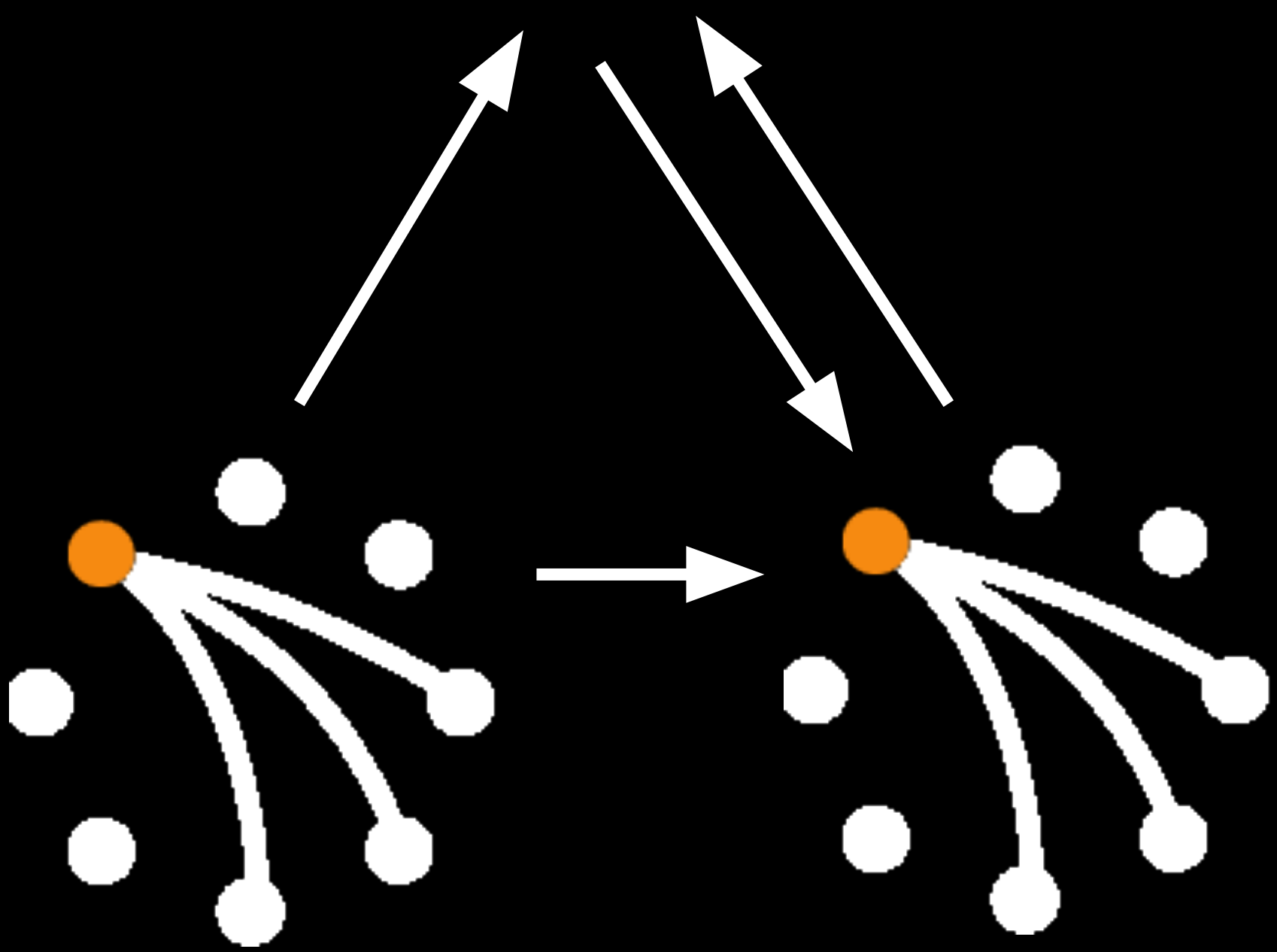
Horrible GET/PUT latency


SUN

Plan B

Abandon ship





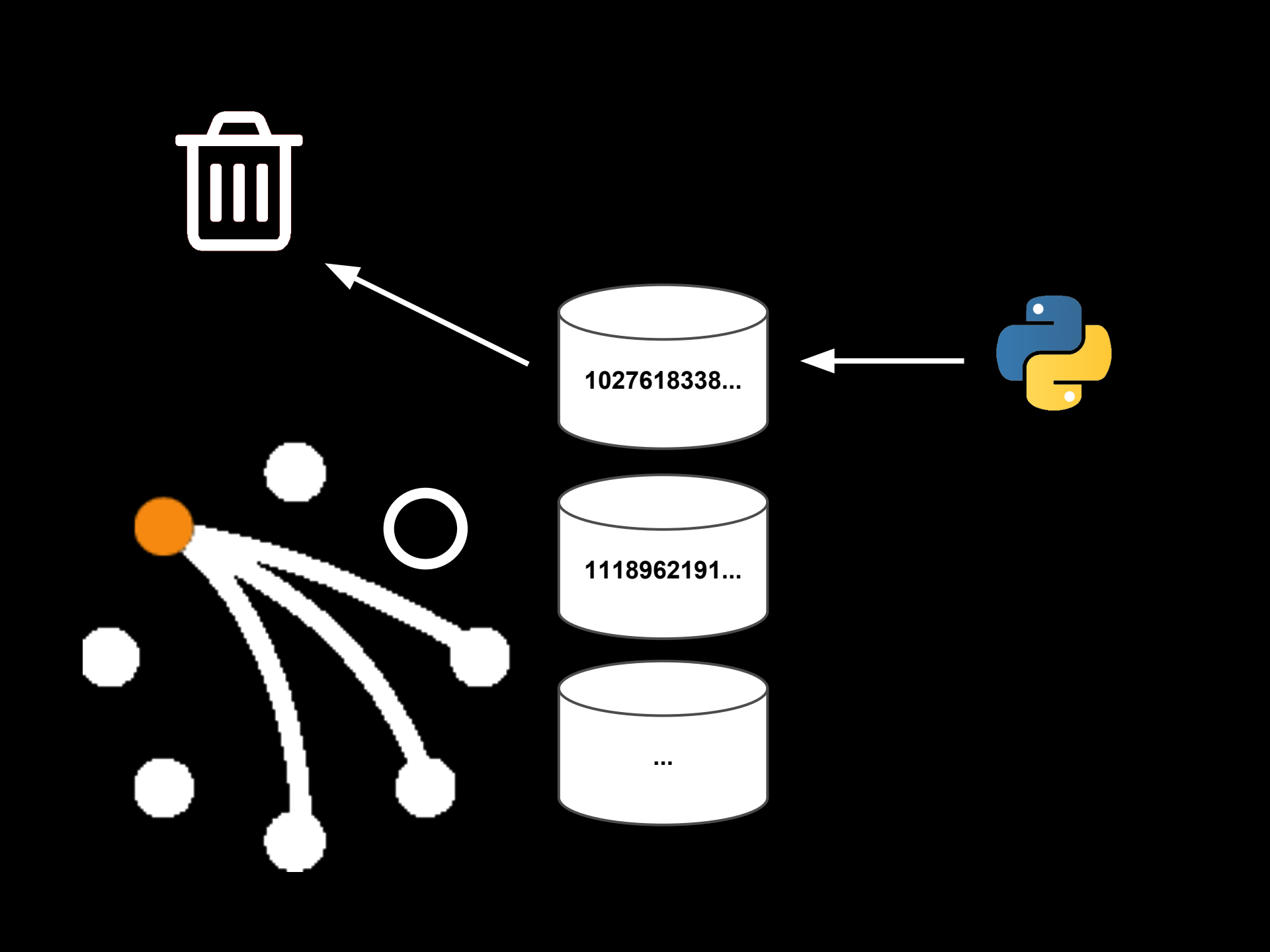
Plan C

Bailout

Need something smarter

Or dumber, dumber is also good!


1027618338...
1118962191...
...


1027618338...
1118962191...
...

It works!

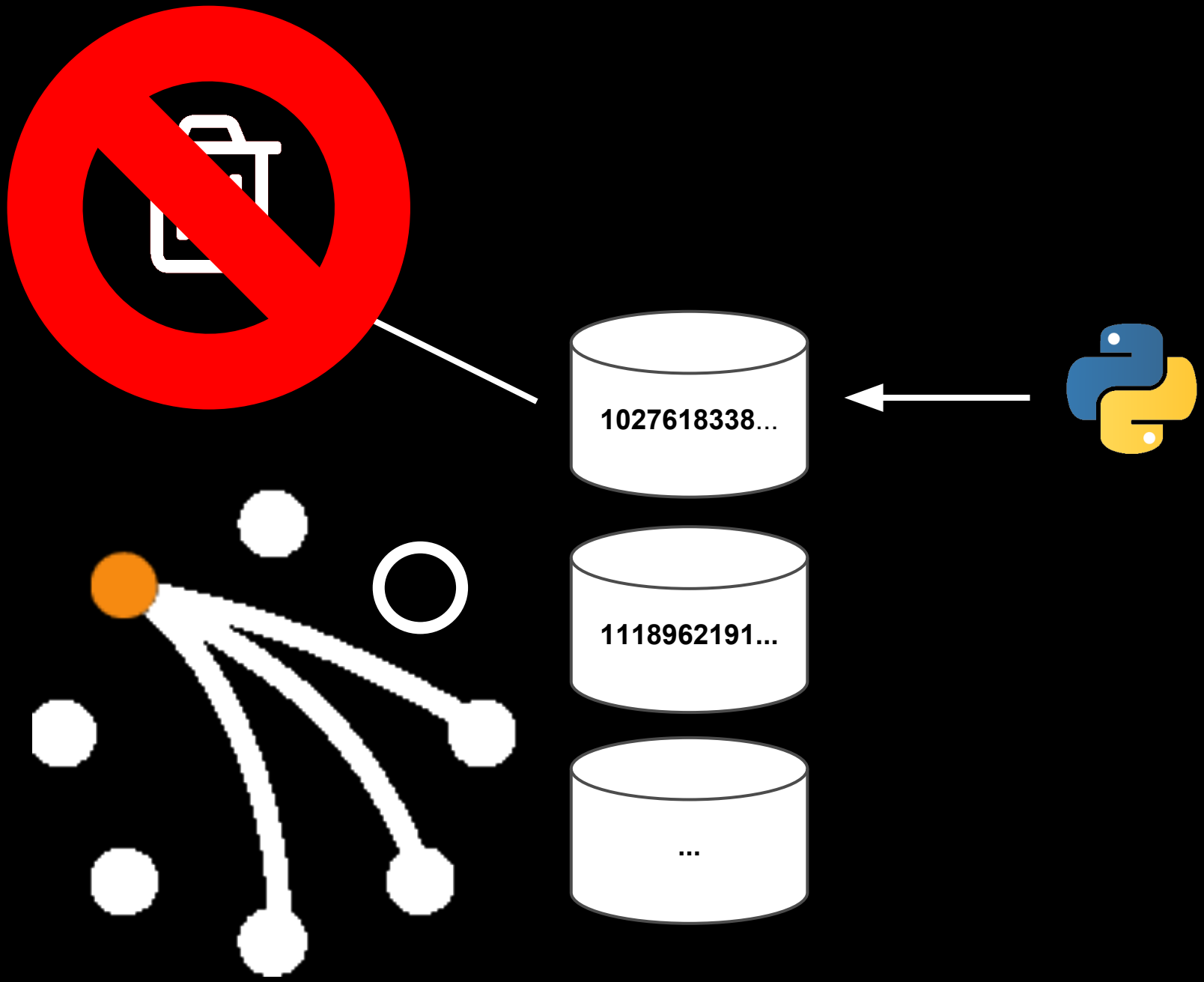
20/sec? wat?

15gb/node/day

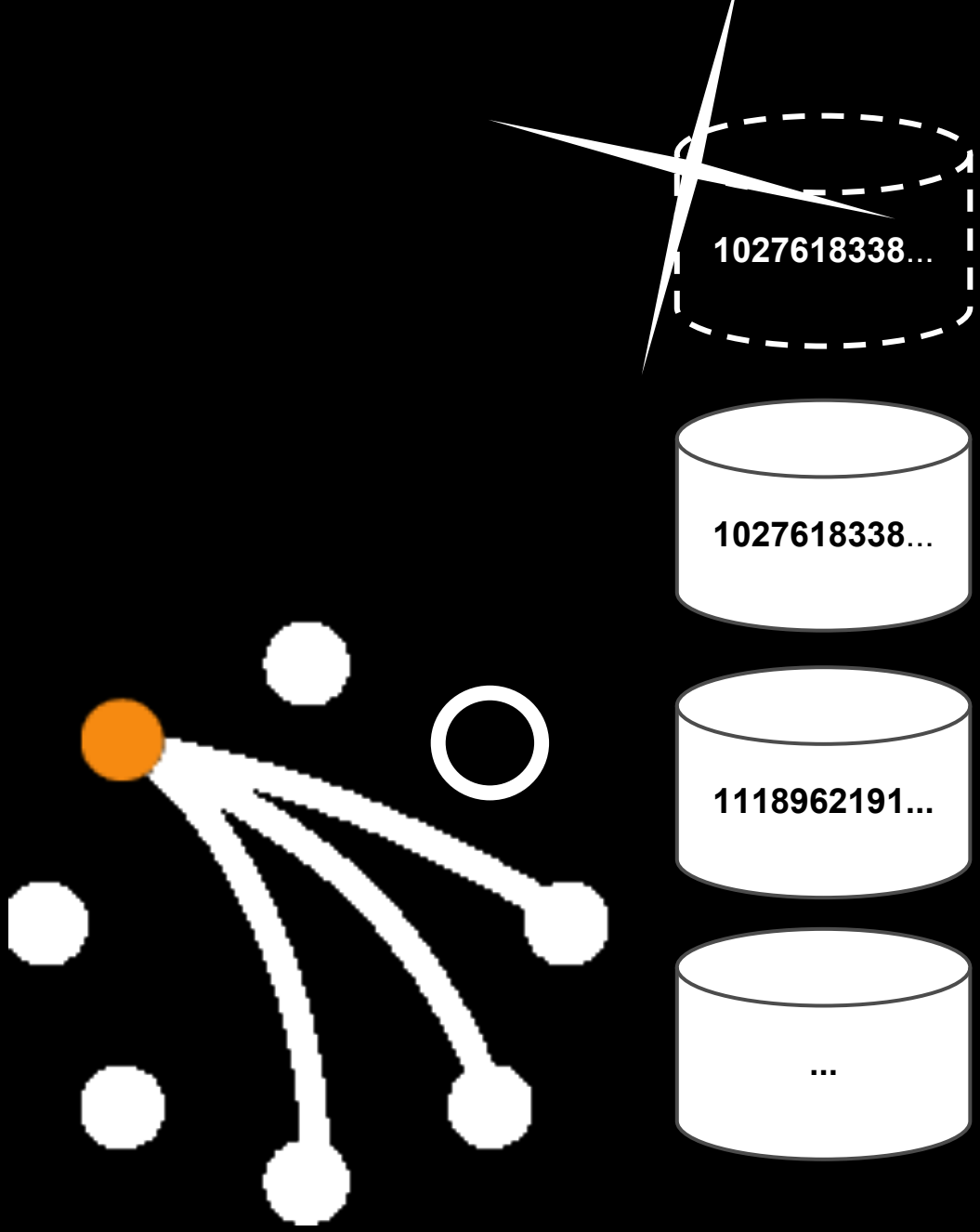
Okay, let's try something really dumb.

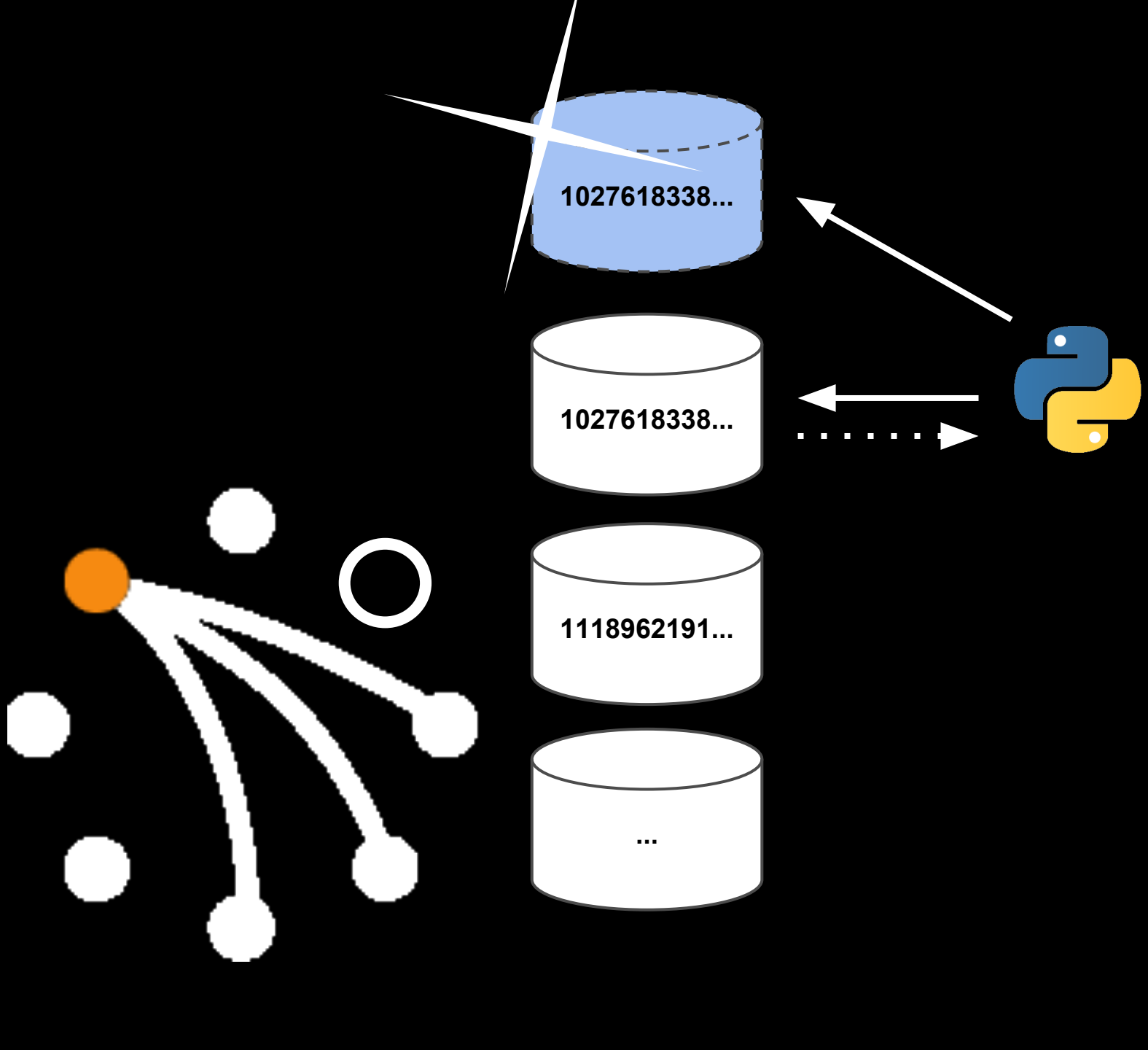
Plan D

Abandon ship... in pieces...


1027618338...
1118962191...
...


1027618338...
1118962191...
...


1027618338...
1027618338...
1118962191...
...


1027618338...
1027618338...
1118962191...
...

2,000 keys/sec!

50% of k/v pairs have no values...

Keys use an unusual serialisation format

Ah, secondary index entries.

Better copy some of those too...

0.5% of keys were kept

Estimated total number of keys: 1.5bn
(without indexes or replication)

Don't do this.

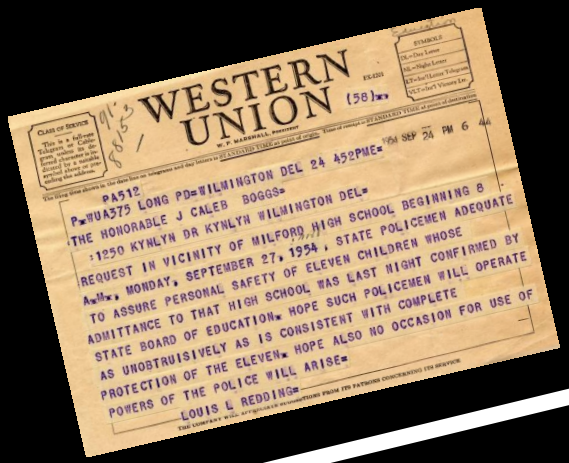
Delete early, delete often.

If you have to do this, do it in Erlang.

sext deserialisation in python

```
<< (<< <<1:1, B1:8>> || <<B1>> <= B >>)/bitstring, 0:Pad, 8 >>.
```

Network latency



WESTERN UNION

W. P. MARSHALL, PRESIDENT

PA512
P=WUA375 LONG PD=WILMINGTON DEL 24 452PME=
THE HONORABLE J CALEB BOGGS=
1250 KYNLYN DR KYNLYN WILMINGTON DEL=
REQUEST IN VICINITY OF MILFORD HIGH SCHOOL BEGINNING 8
A=M=, MONDAY, SEPTEMBER 27, 1954, STATE POLICEMEN ADEQUATE
TO ASSURE PERSONAL SAFETY OF ELEVEN CHILDREN WHOSE
ADMITTANCE TO THAT HIGH SCHOOL WAS LAST NIGHT CONFIRMED BY
STATE BOARD OF EDUCATION= HOPE SUCH POLICEMEN WILL OPERATE
AS UNOBTRUISIVELY AS IS CONSISTENT WITH COMPLETE
PROTECTION OF THE ELEVEN= HOPE ALSO NO OCCASION FOR USE OF
POWERS OF THE POLICE WILL ARISE=
LOUIS L REDDING=

100ms

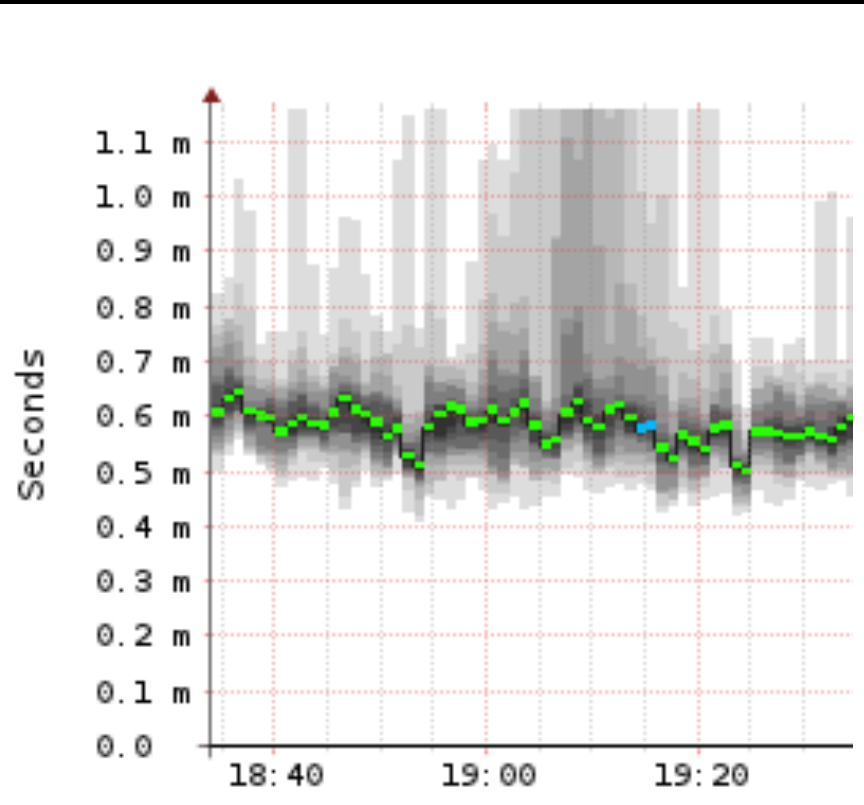
Using a peer to peer VPN

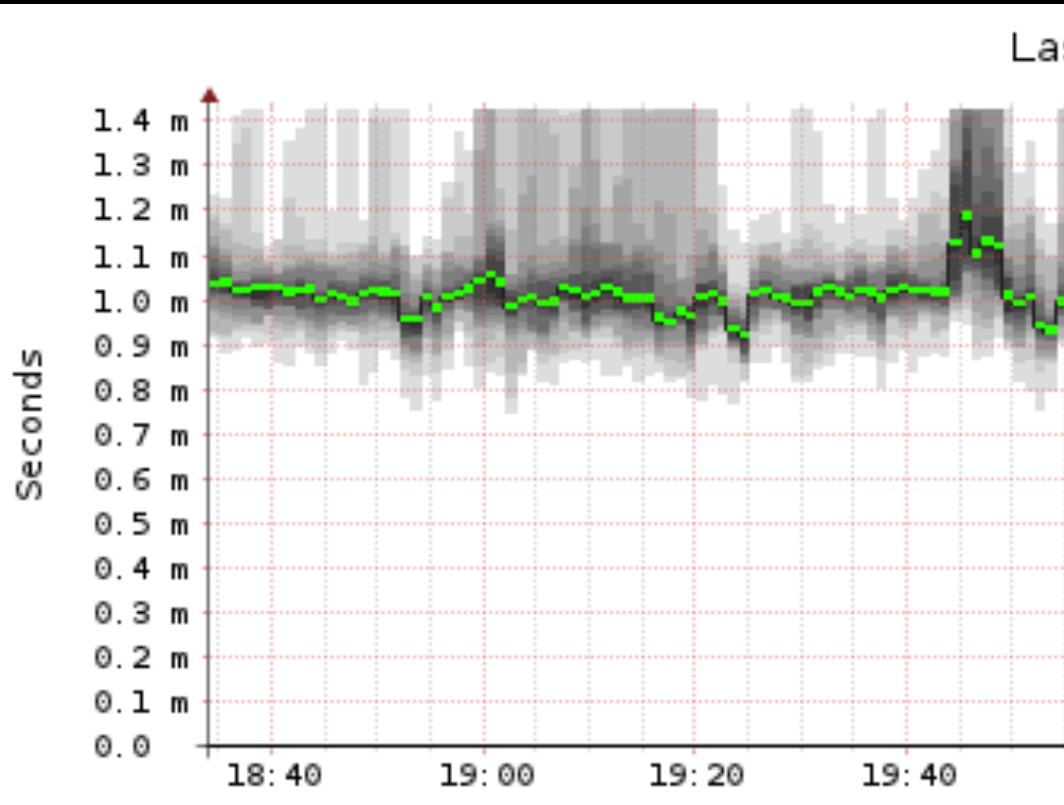
Wasn't very p2p

Let's just go off to this other corner of the internet.

p2p VPN configured correctly

Adds 500µs, just about acceptable





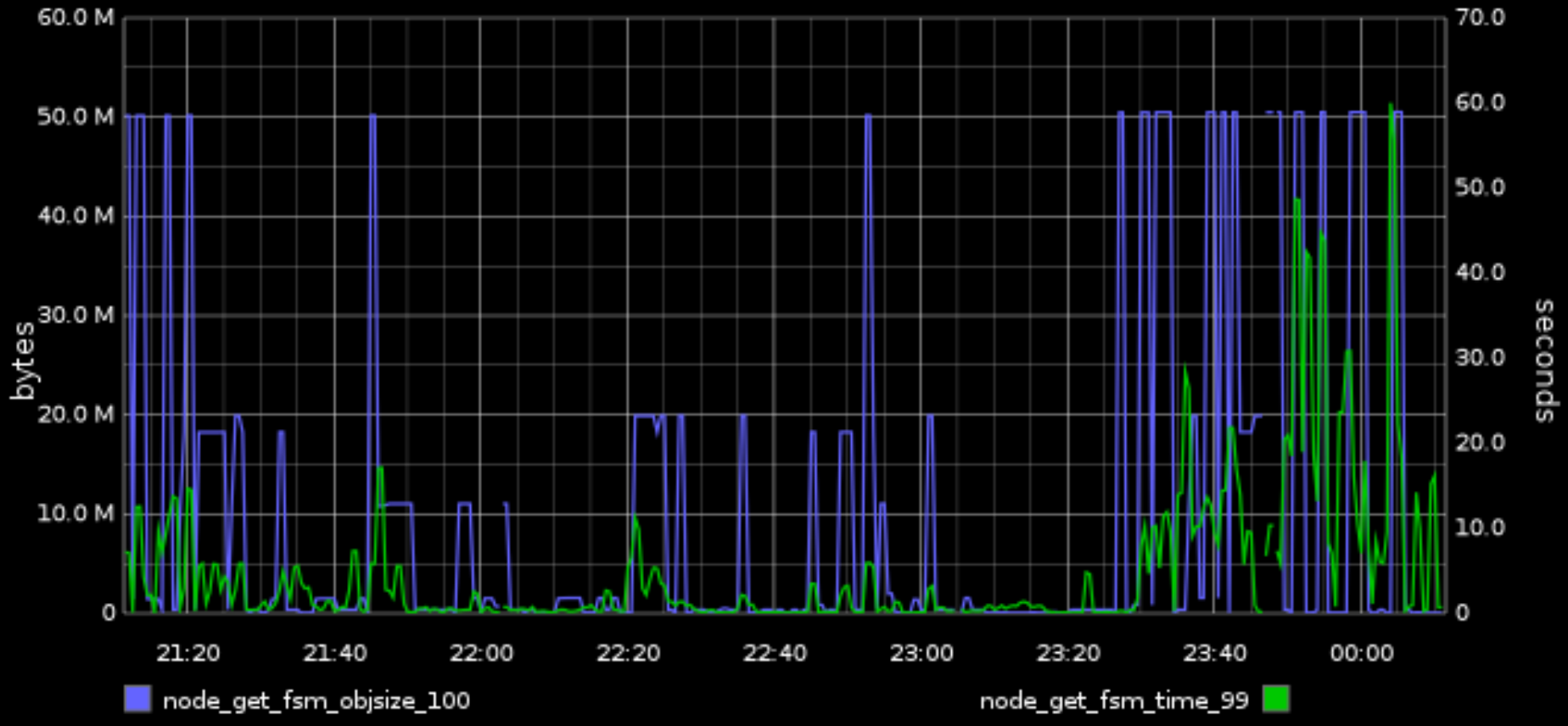
Maximum object sizes

5mb, they said.

Accidentally created a 50mb object

... and growing every few minutes

What a 50mb object does to your response times



Timeout enforced by haproxy in front of riak

Don't do that.

Siblings

Siblings are normal

Created and resolved all the time, intentionally

6,000 siblings

10k object, 6,000 copies... oh.

Don't do that.

Resolve siblings everywhere.

Recommendations

Things I can't recommend

"None of yer fancy stuff"

Using mapreduce for real-timey work

It sure does look good in dev though.

Using 2i for real-timey work

It sure does look good in dev though.

Running other services

OOM killed, and slow nodes worse than down node.

Unpredictable performance on EC2

Haven't done it myself...

Things I can recommend

This shit mostly just works.

Node adding, removing, failure and recovery

Straight key-value GET/PUT

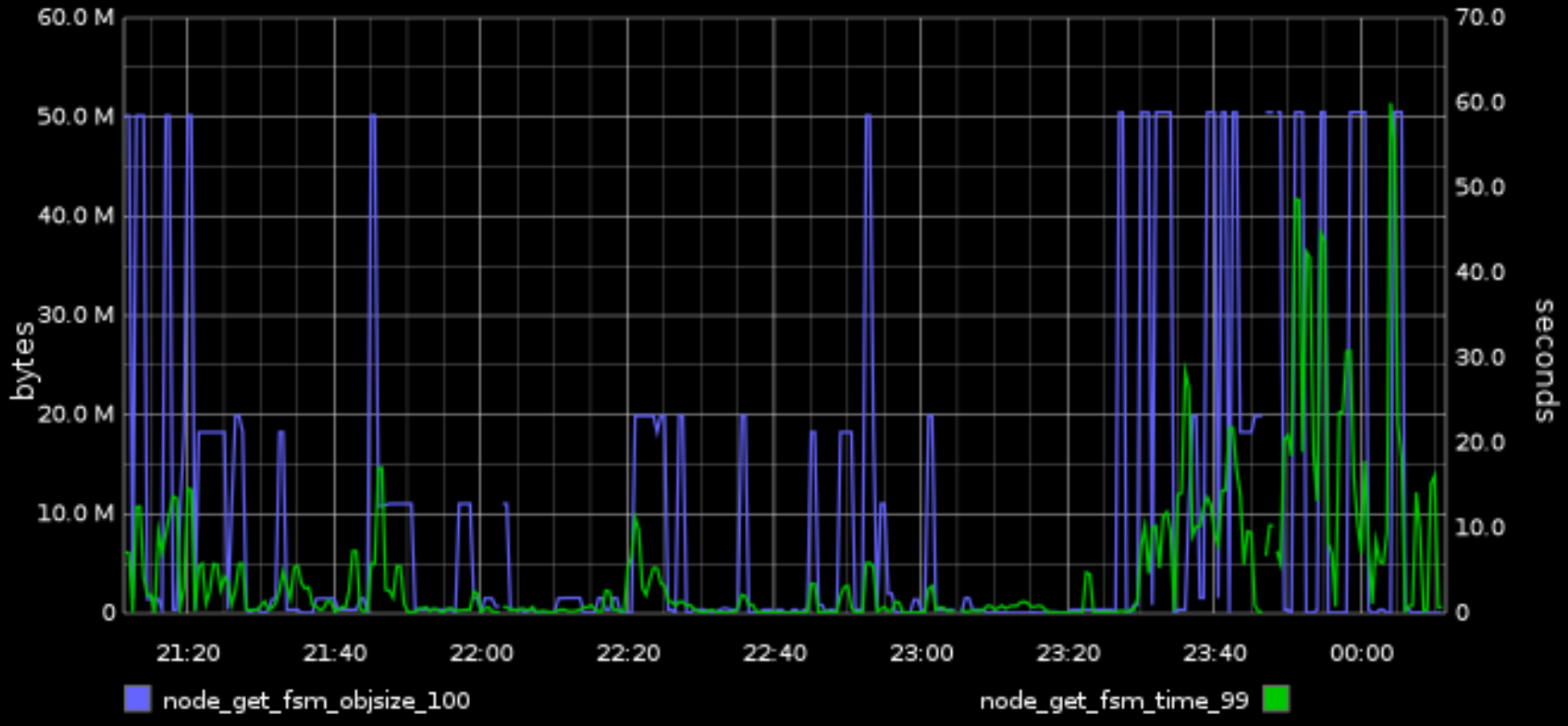
Siblings and sibling resolution

Grumpy recommendations

COLLECT STATISTICS

When your cluster is broken, you won't have ANY IDEA WHY!

What a 50mb object does to your response times



COLLECT STATISTICS

But of course a stats company would say that...

HTTP vs protobuf

Use predictable keys

Generate key names, don't search for them.

Don't touch ring_creation_size

Unless you know better, the default is just fine!

Run five nodes!

On distinct hardware!

Use the community!

IRC: #riak on Freenode

Do what Basho tell you.

Riak probably won't wake you up.

Hosted Graphite

You send us numbers, we give you graphs.

We're almost hiring!

Looking for engineer #0 to help us abuse Riak.