from
Buoyant

QConSF, November 9, 2016

Freeing the Whale

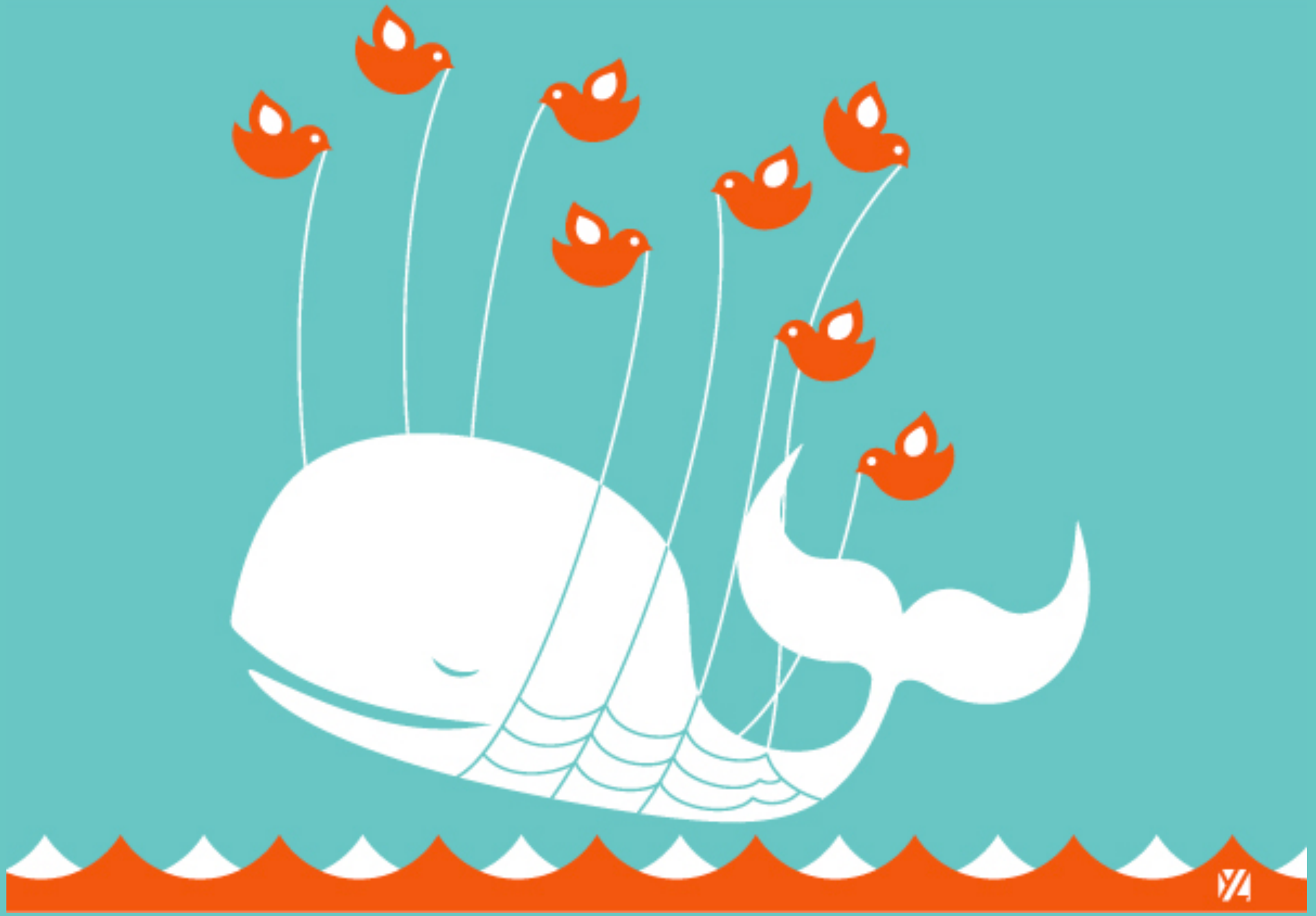
How to Fail at Scale

oliver gould
cto, buoyant

2010

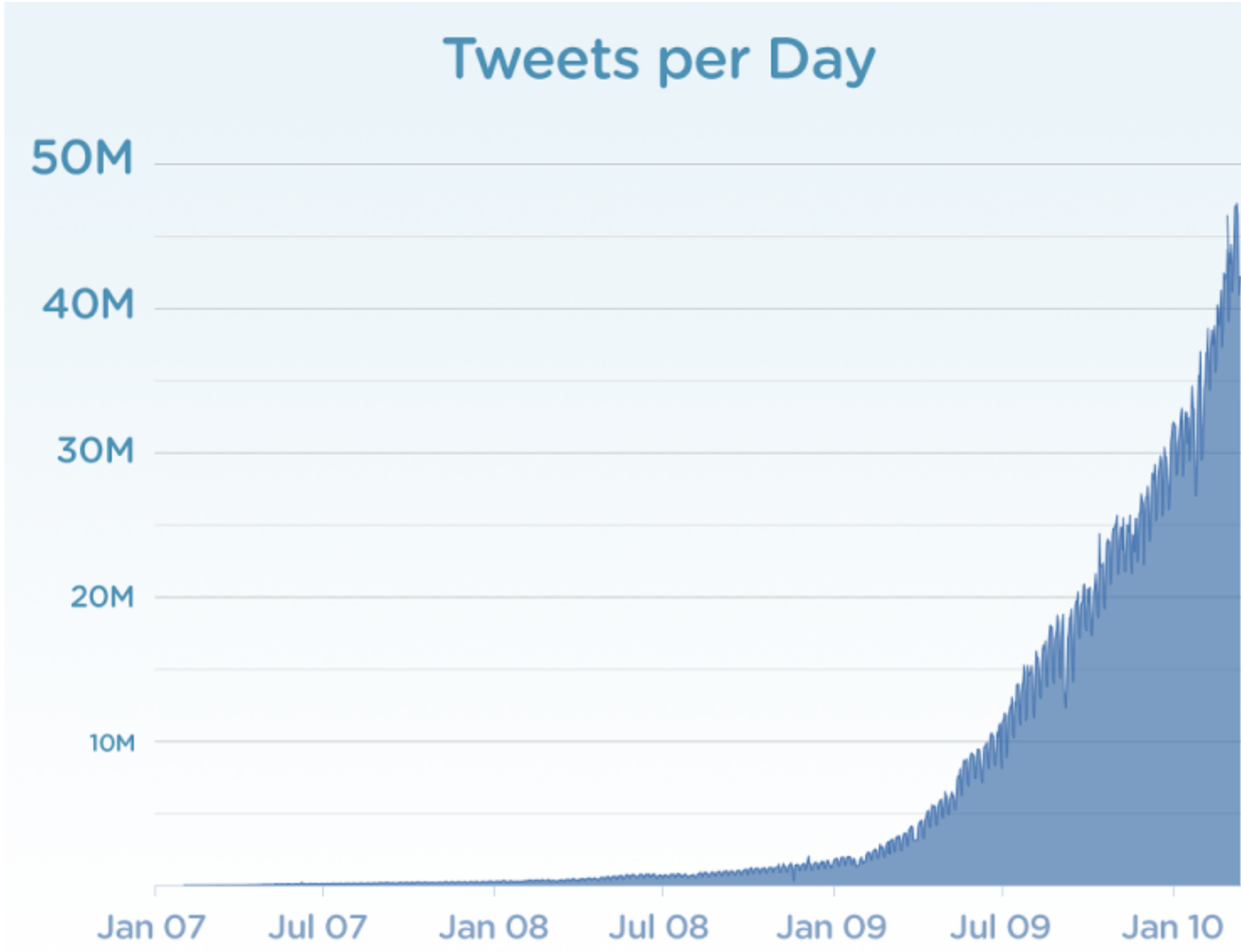
A FAILWHALE ODYSSEY

2010 FIFA World Cup



Twitter, 2010

- 10^7 users
- 10^7 tweets/day
- 10^2 engineers
- 10^1 ops eng
- 10^1 services
- 10^1 deploys/week
- 10^2 hosts
- 0 datacenters
- **10^1 user-facing outages/week**



https://blog.twitter.com/2010/measuring-tweets

Source: twitte

objective

reliability

flexibility

objective	solution
reliability	platform
flexibility	SOA + devops i.e. “microservices”

Resilience is an imperative: our software runs on the truly dismal computers we call *datacenters*. Besides being heinously complex… they are unreliable and prone to operator error.

Marius Eriksen
@marius
RPC Redux

software you didn't write

hardware you can't touch

network you can't trace

break in new and surprising ways

and **your customers** shouldn't notice

freeing the whale

photo: Johanan Ottensooser

mesos.apache.org

UC Berkeley, 2010

Twitter, 2011

Apache, 2012

Abstracts compute resources

Promise: don't worry about the hosts

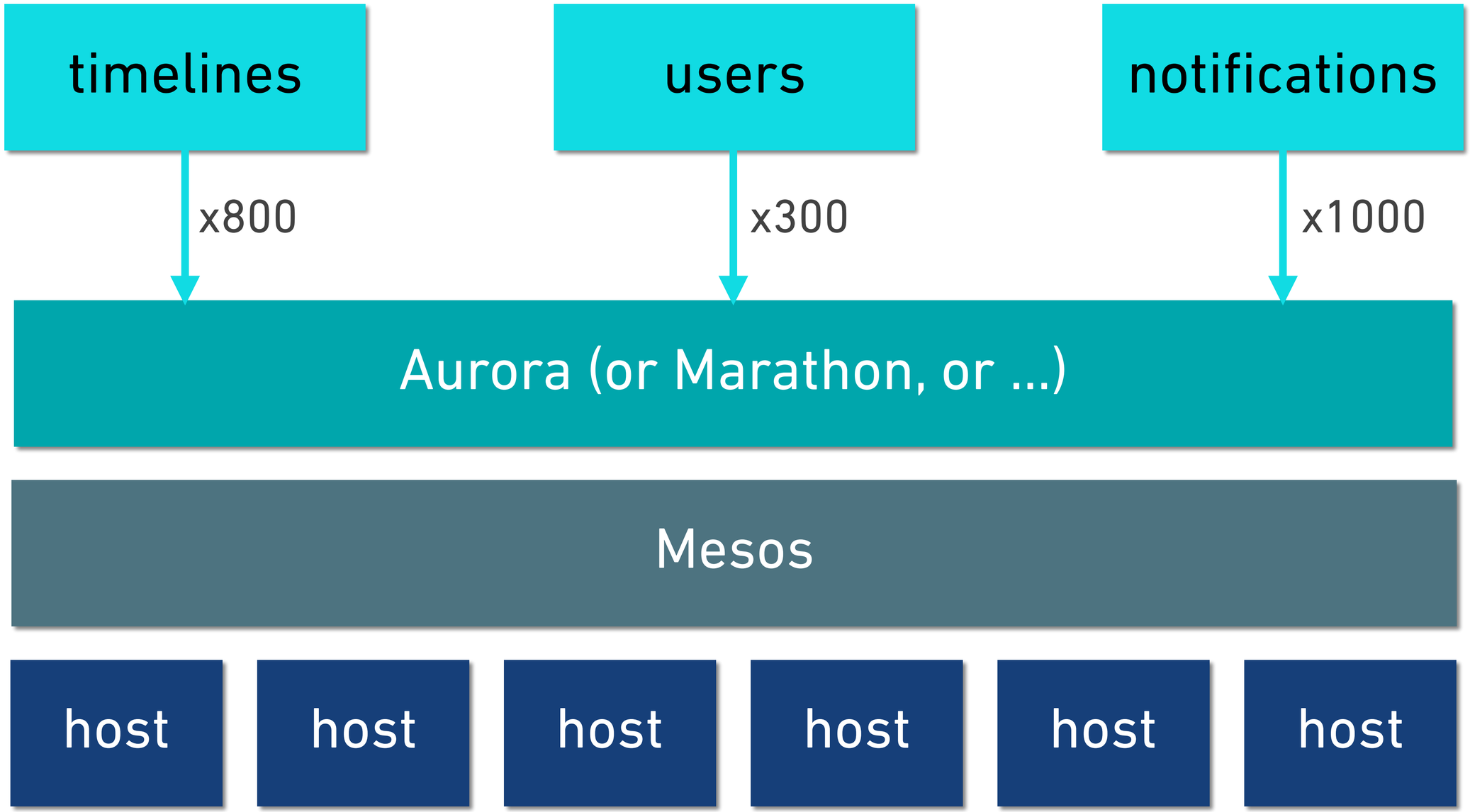
aurora.apache.org

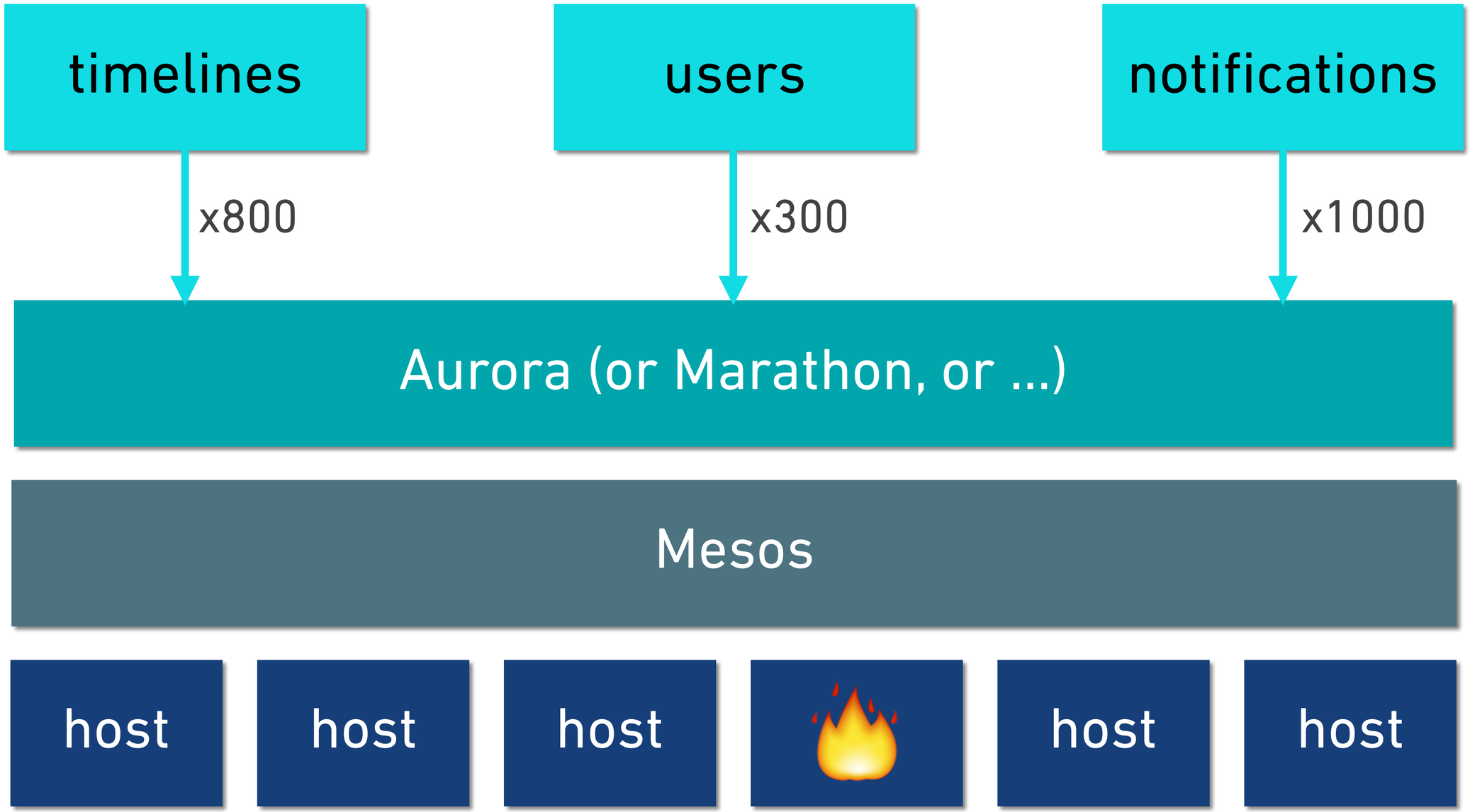
Twitter, 2011

Apache, 2013

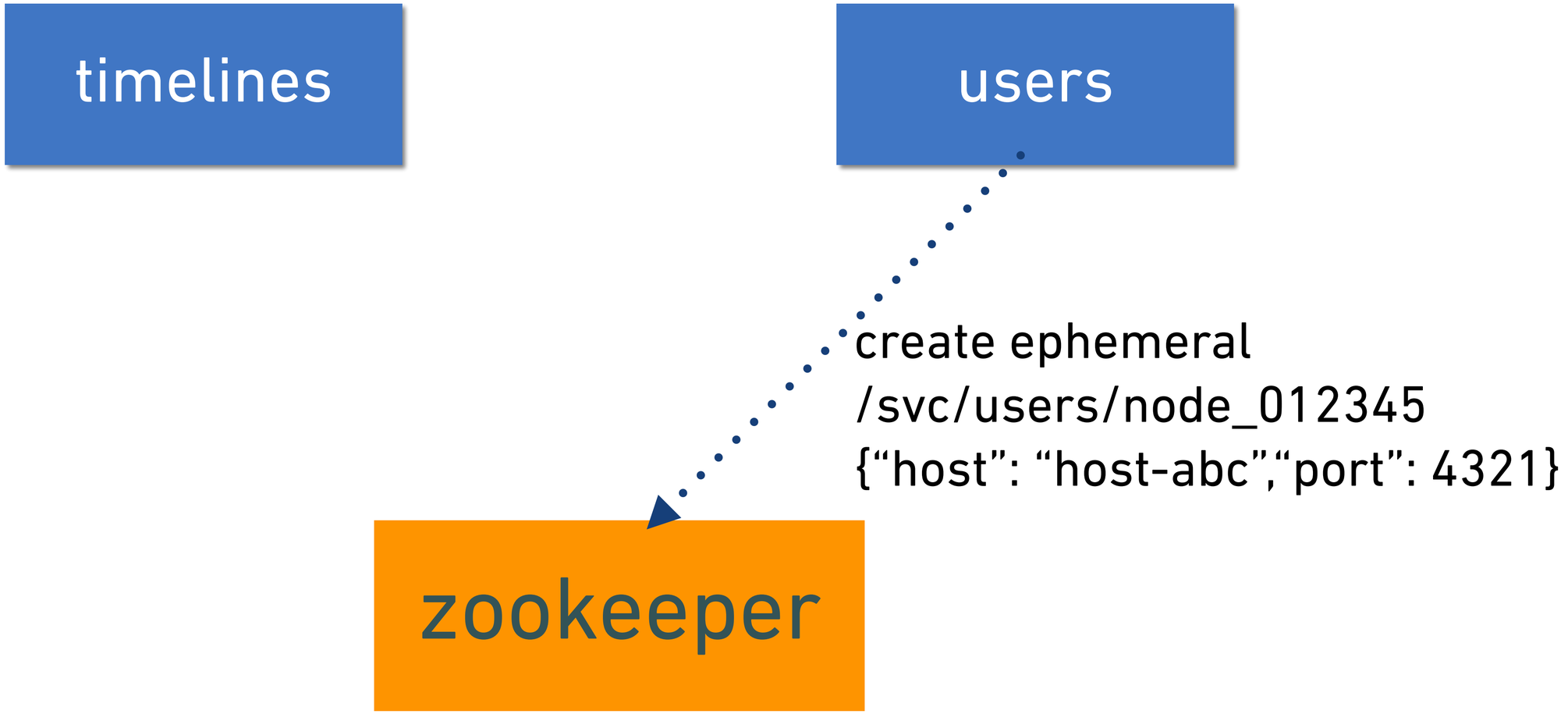
Schedules processes on Mesos

Promise: no more puppet, monit, etc

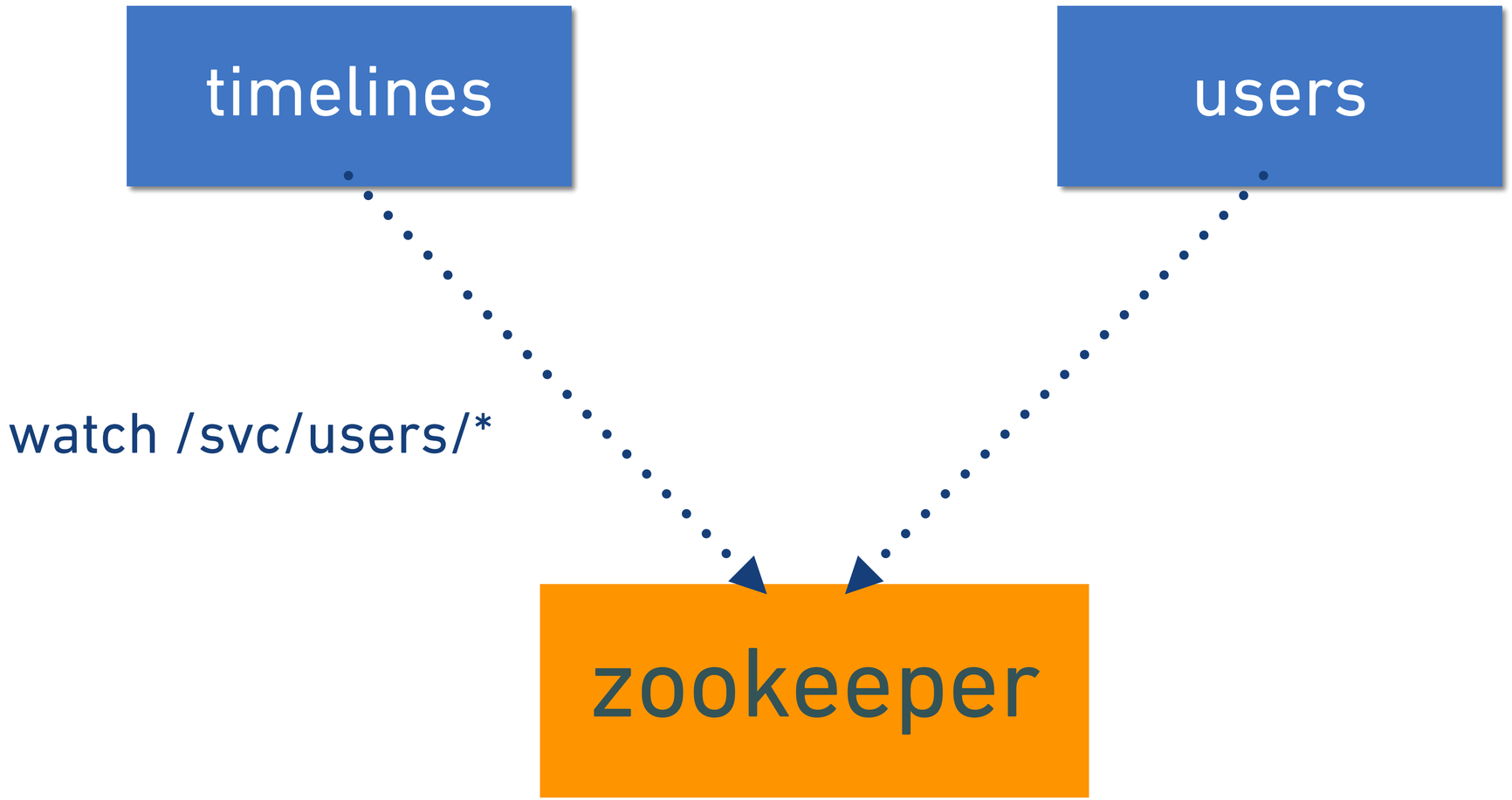

timelines
users
notifications
x800
x300
x1000
Aurora (or Marathon, or ...)
Mesos
host
host
host
host
host
host


timelines
users
notifications
x800
x300
x1000
Aurora (or Marathon, or ...)
Mesos
host
host
host
host
host

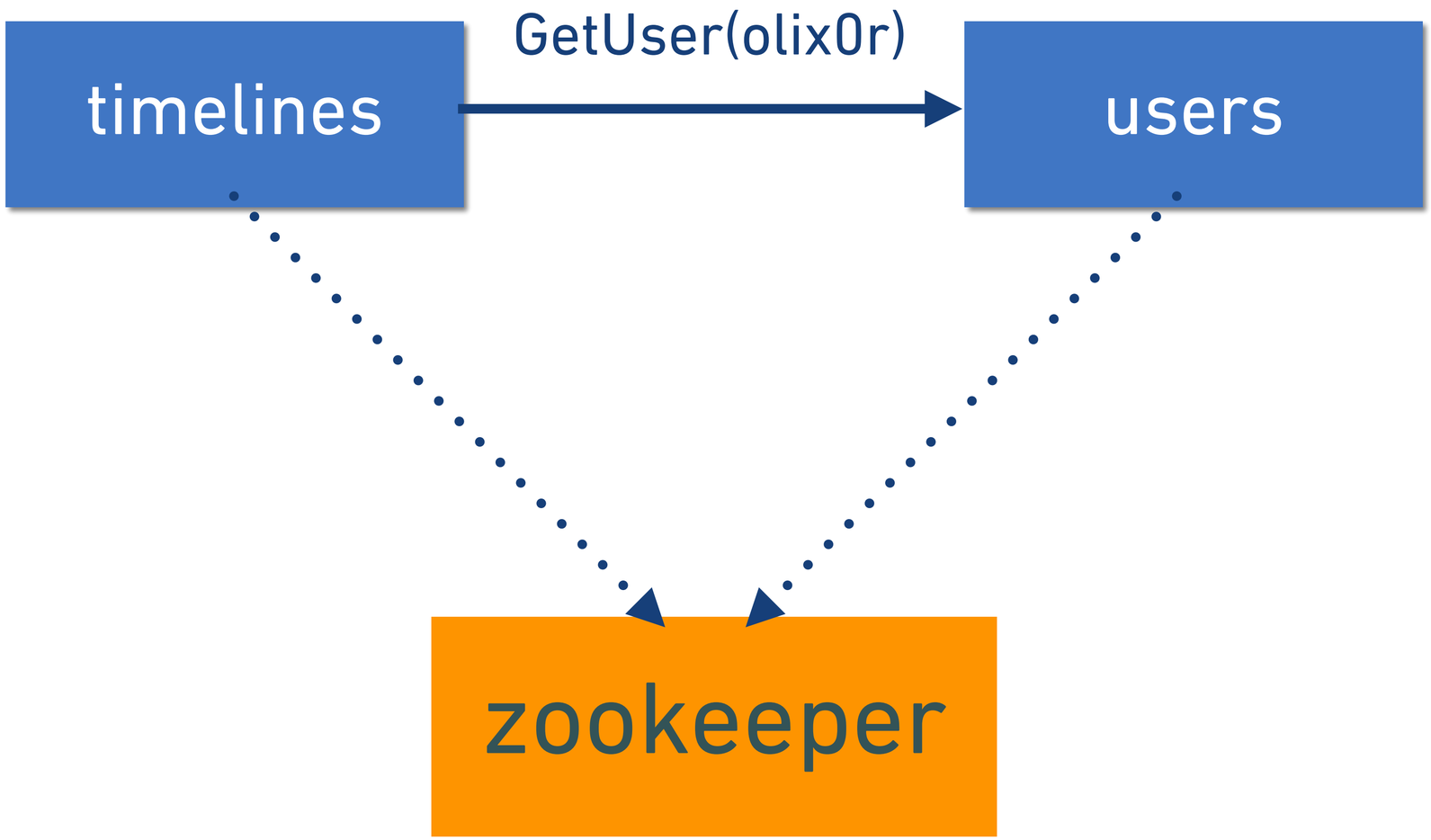
service discovery



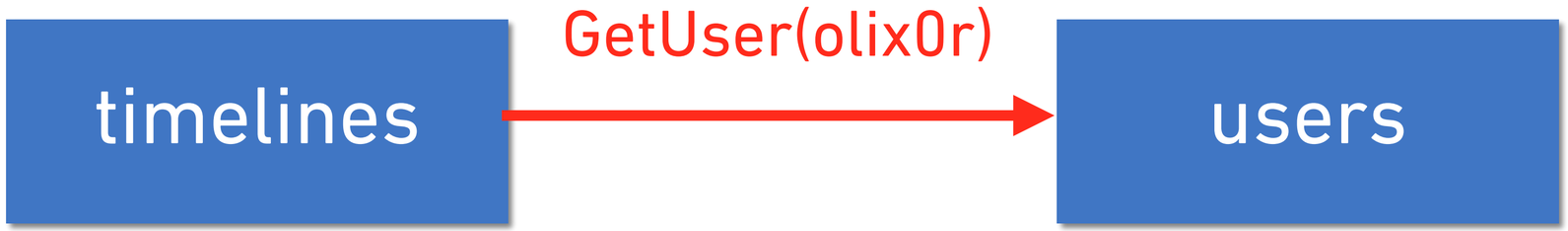
service discovery



service discovery



service discovery



uh oh.

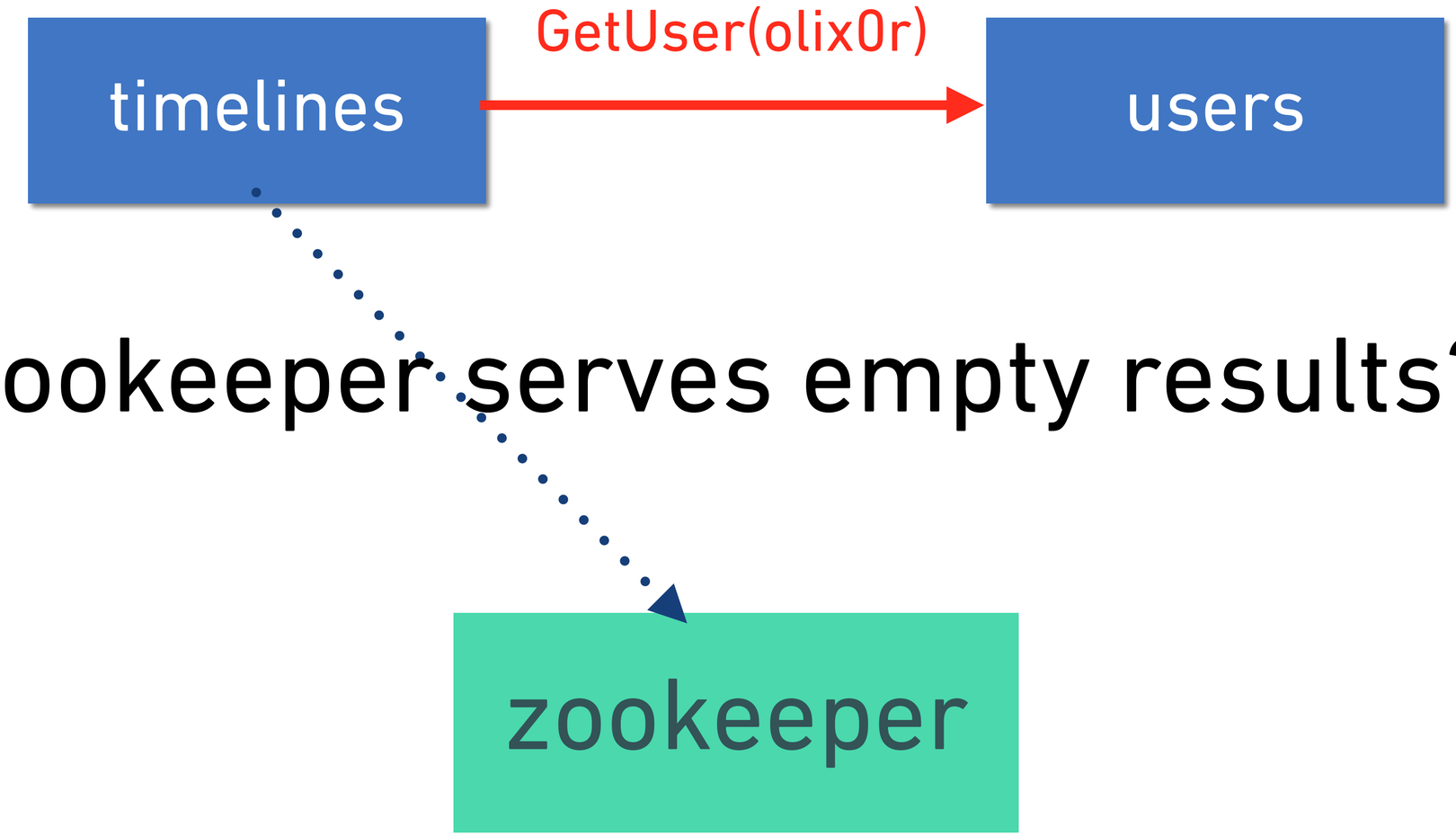
zookeeper

service discovery

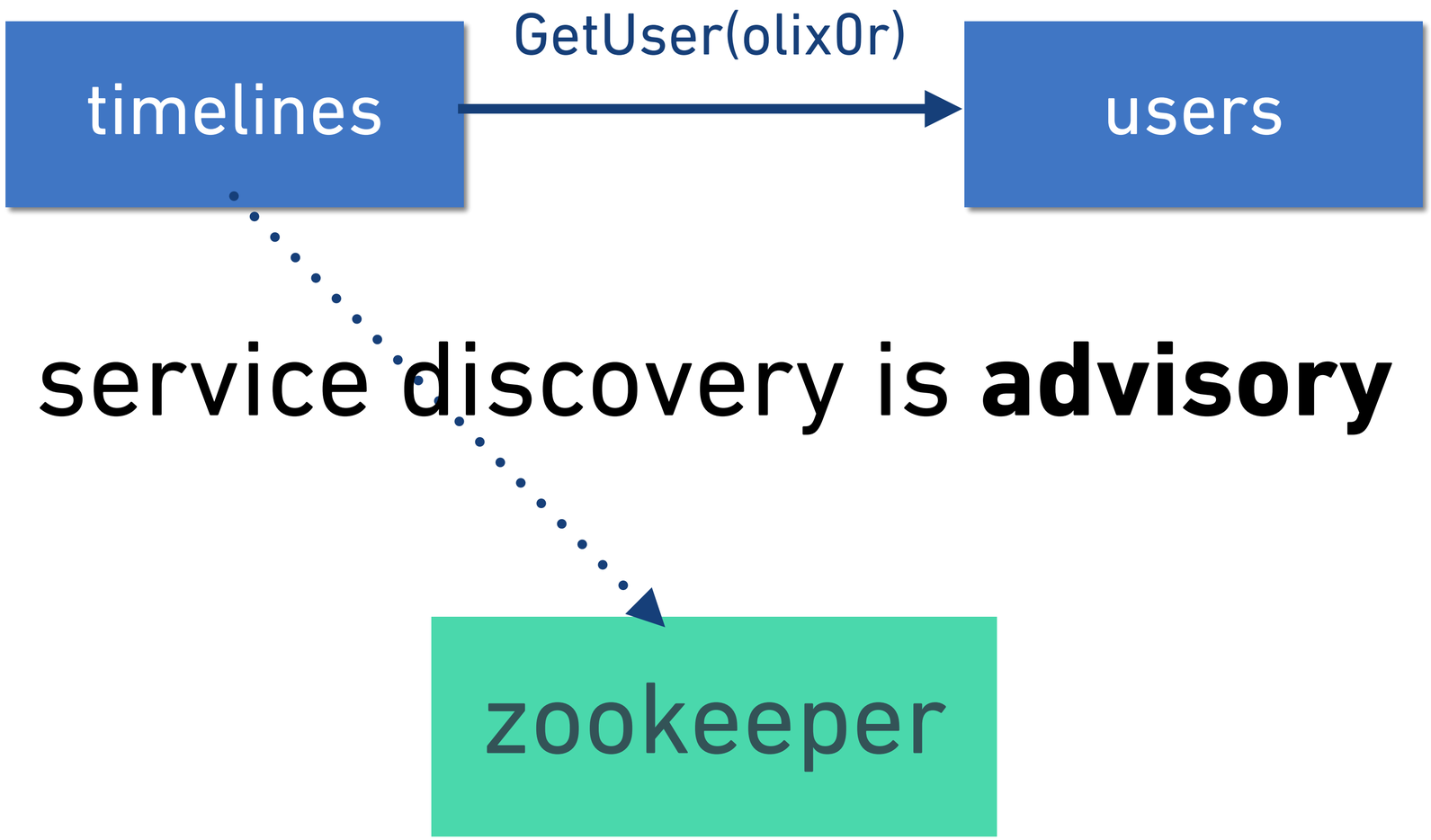
client caches results

zookeeper

service discovery

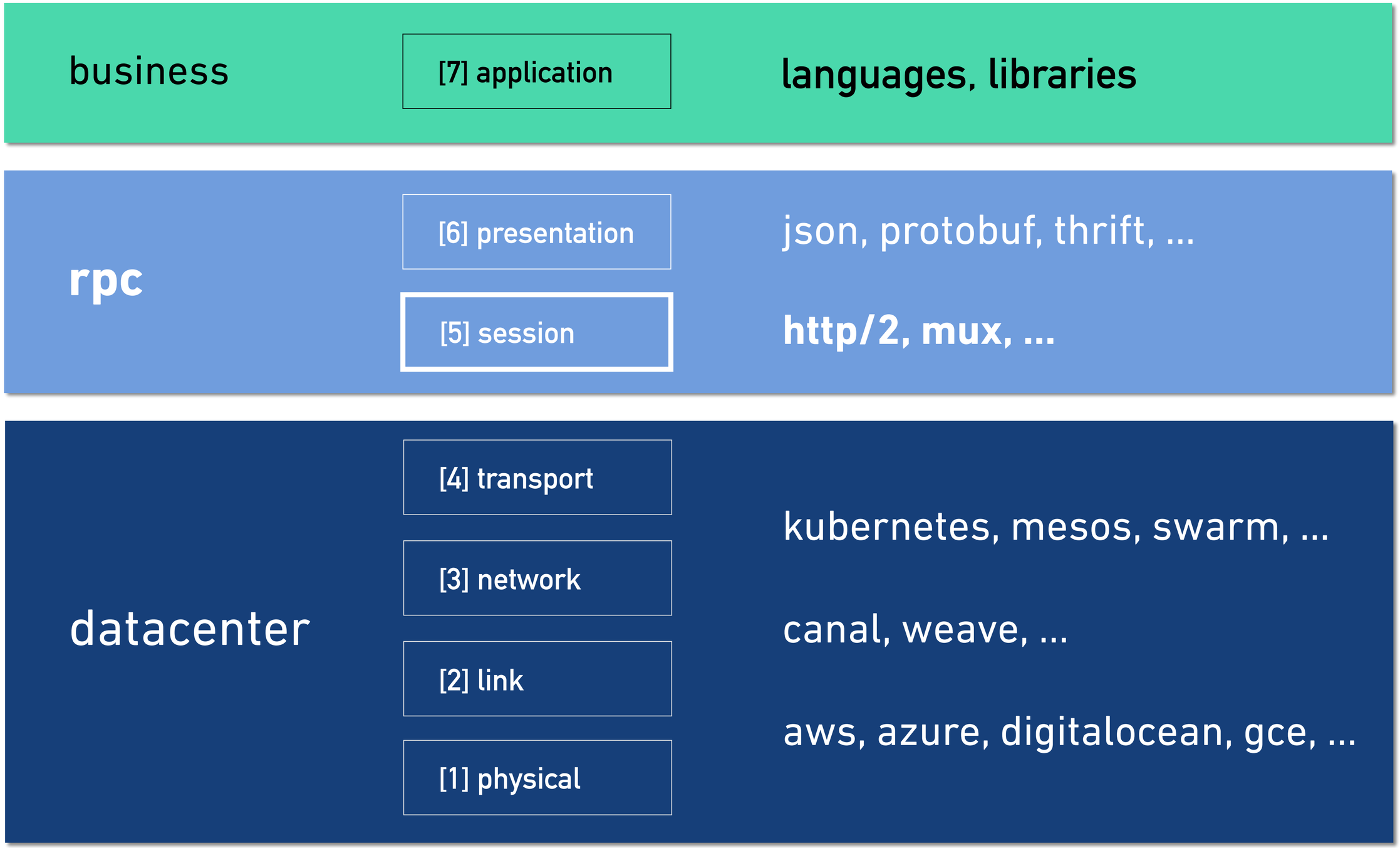


service discovery



github.com/twitter/finagle

- RPC **library** (JVM)
- asynchronous
- built on **Netty**
- scala
  - functional
  - strongly typed
- first commit: Oct 2010


business
[7] application
languages, libraries
rpc
[6] presentation
[5] session
json, protobuf, thrift, ...
http/2, mux, ...
datacenter
[4] transport
[3] network
[2] link
[1] physical
kubernetes, mesos, swarm, ...
canal, weave, ...
aws, azure, digitalocean, gce, ...

“**It’s slow**”
is the hardest problem you’ll ever debug.

Jeff Hodges
@jmhodges
Notes on Distributed Systems for Young Bloods

observability

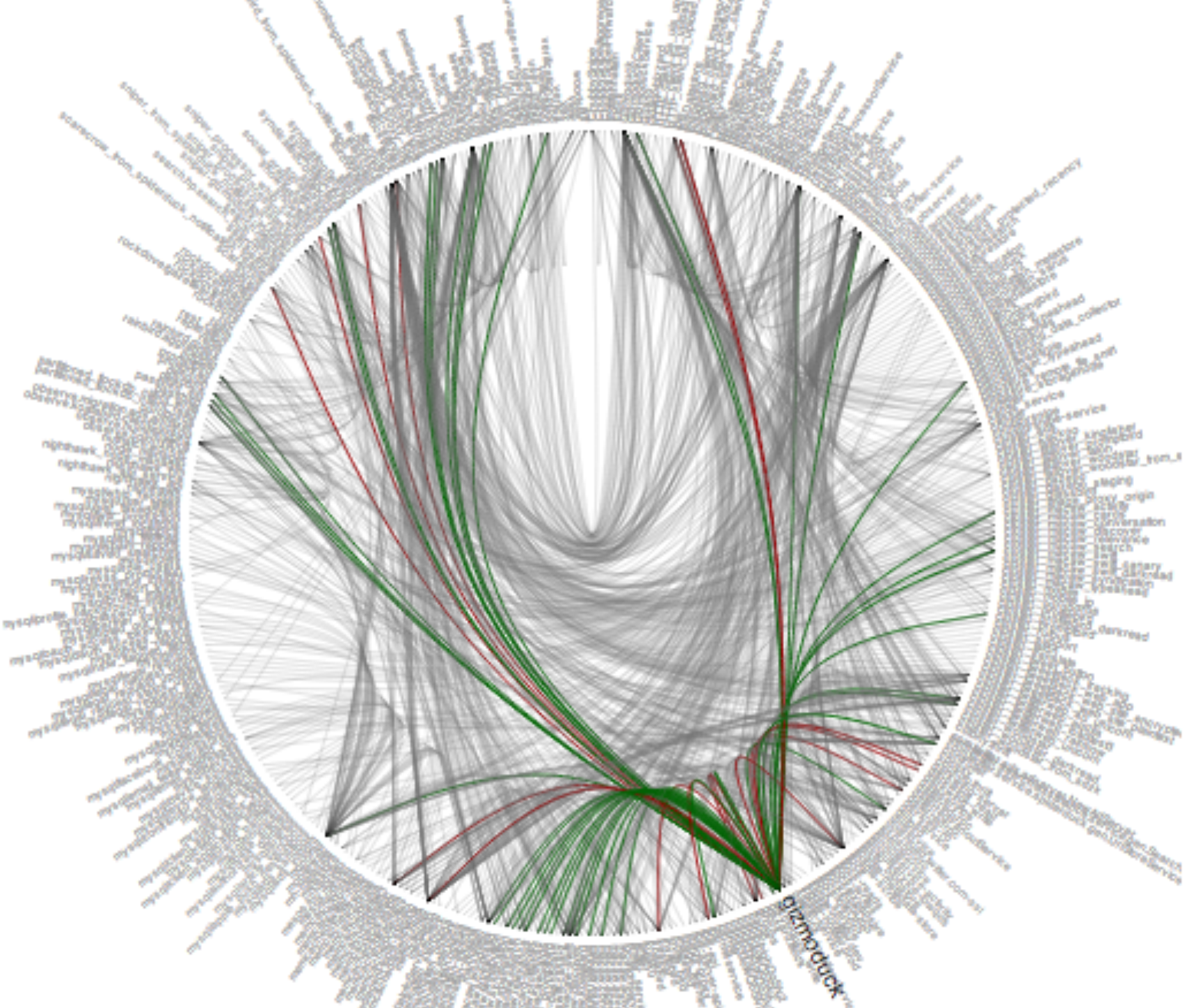
counters (e.g. client/users/failures)

histograms (e.g. client/users/latency/p99)

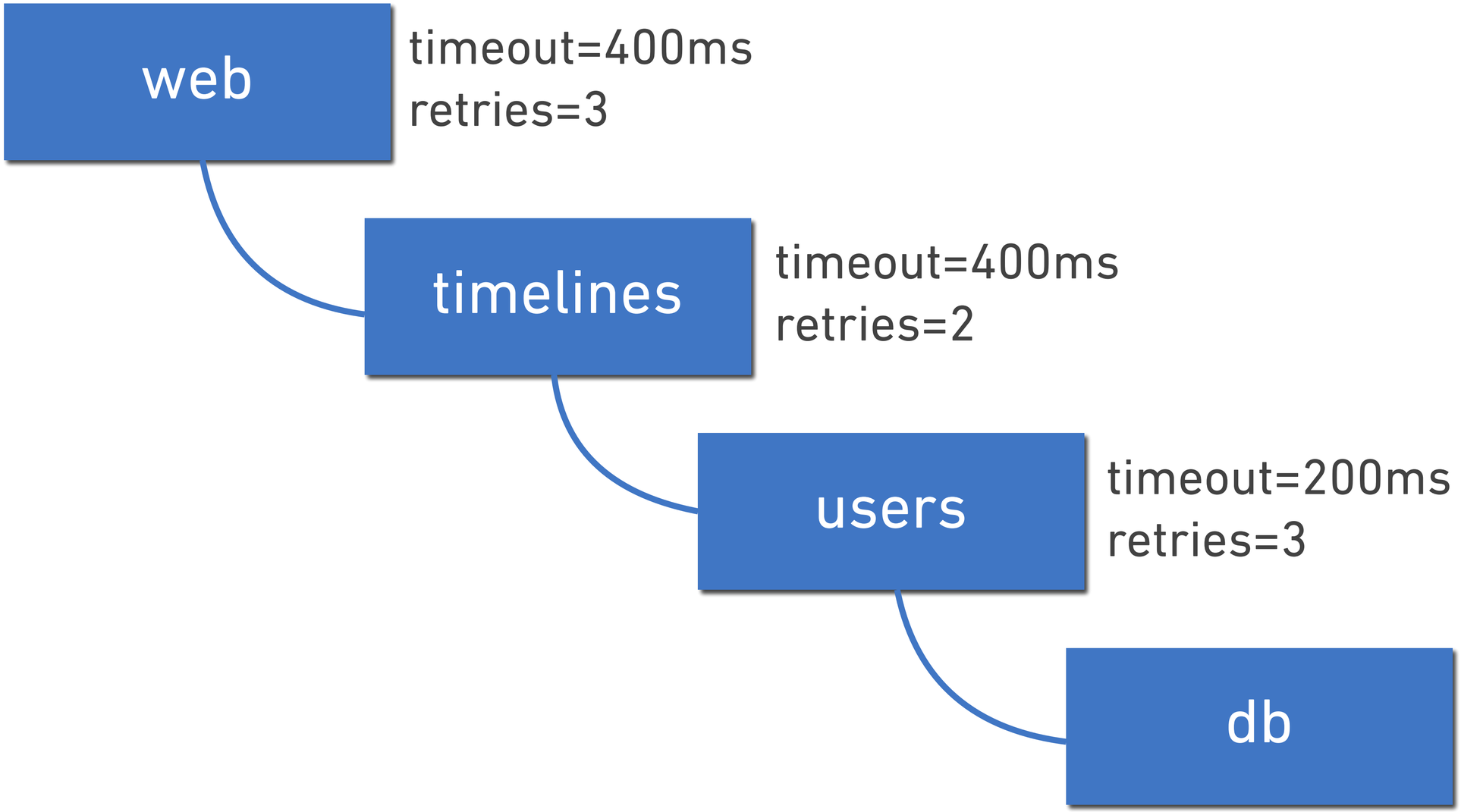
tracing

tracing

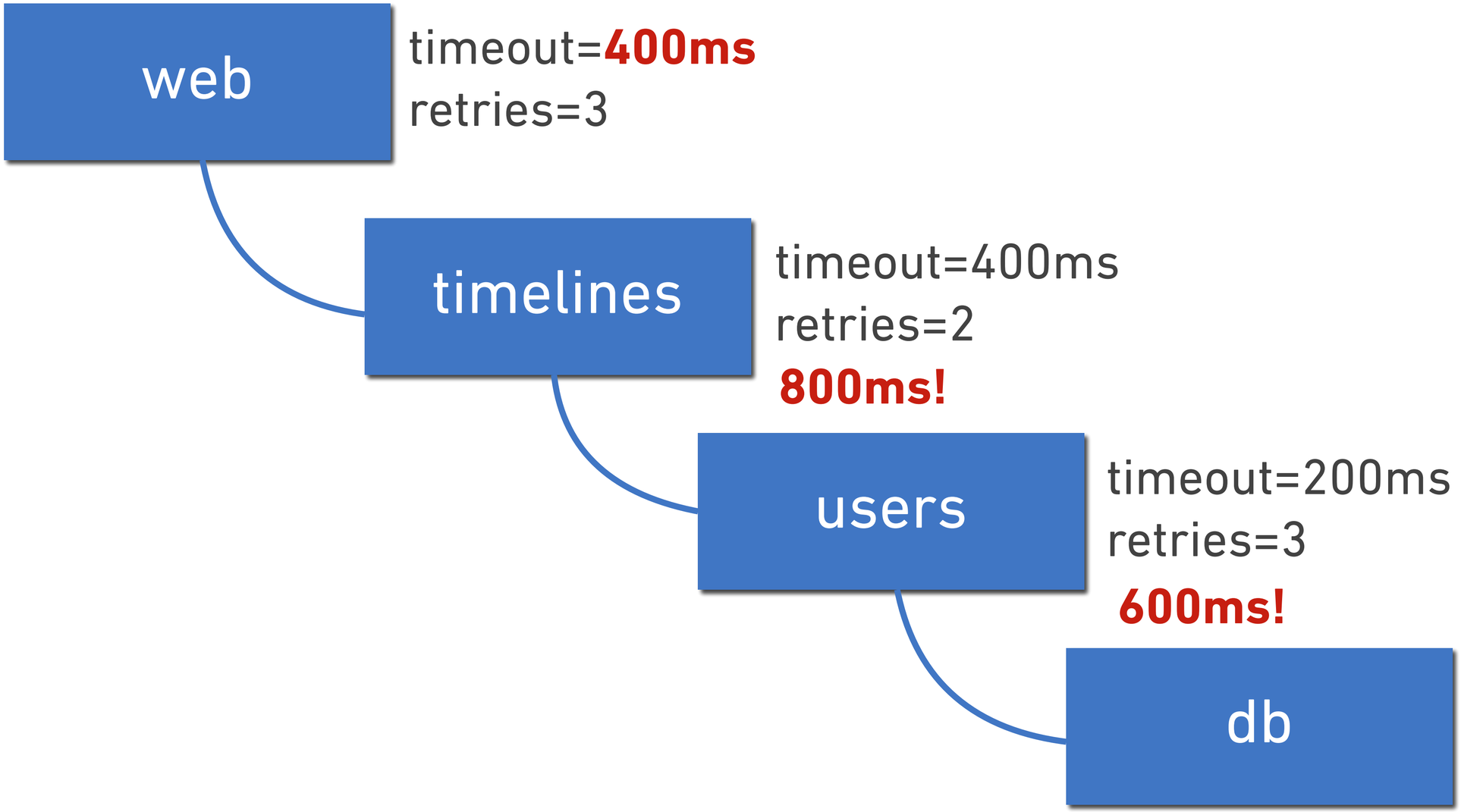
ZIPKIN



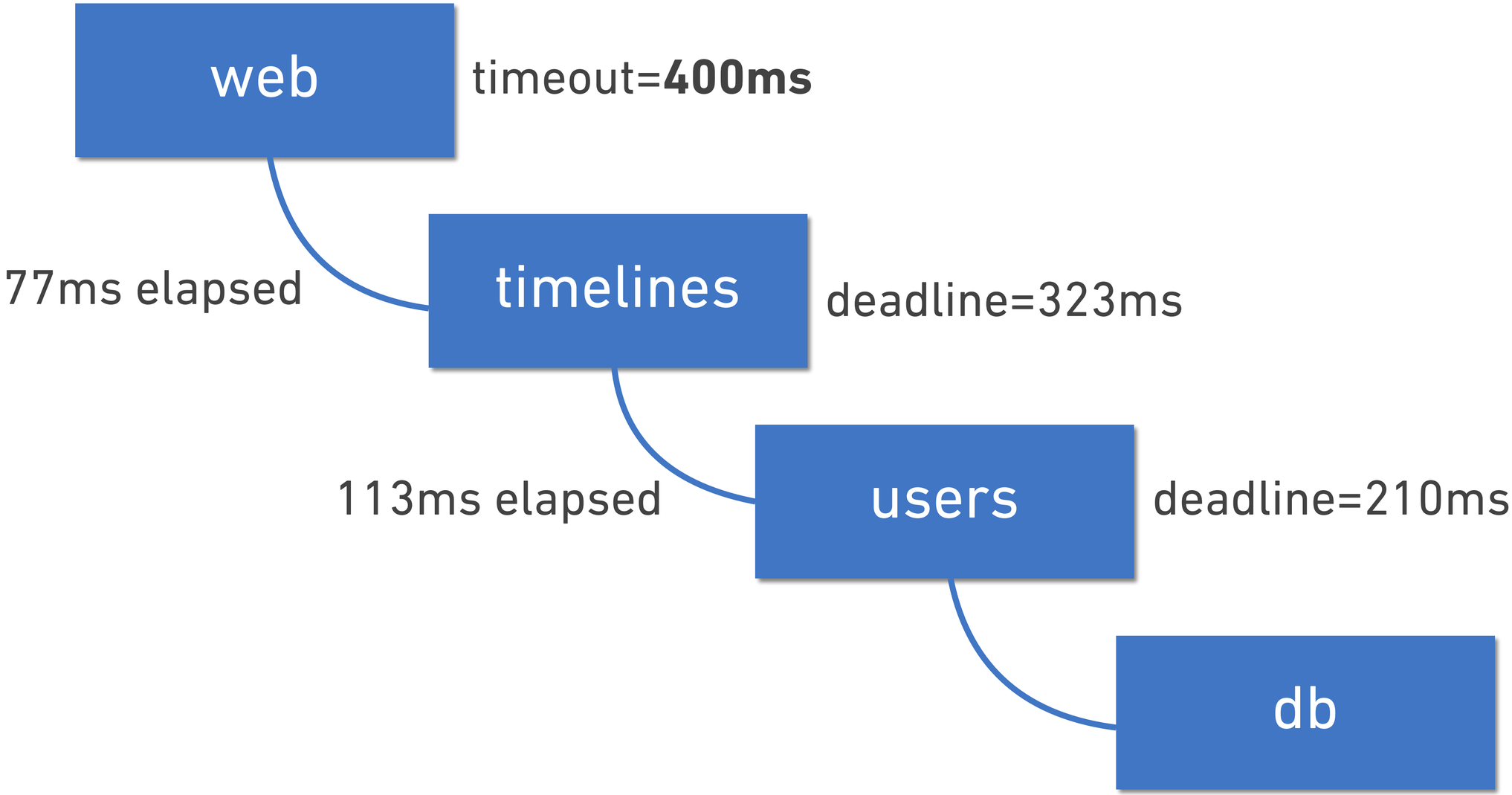
timeouts & retries



timeouts & retries



deadlines


web
timeout=**400ms**
77ms elapsed
timelines
deadline=323ms
113ms elapsed
users
deadline=210ms
db

retries

typical:

retries=3

retries

typical:

retries=3

worst-case: 300% more load!!!

budgets

typical:

retries=3

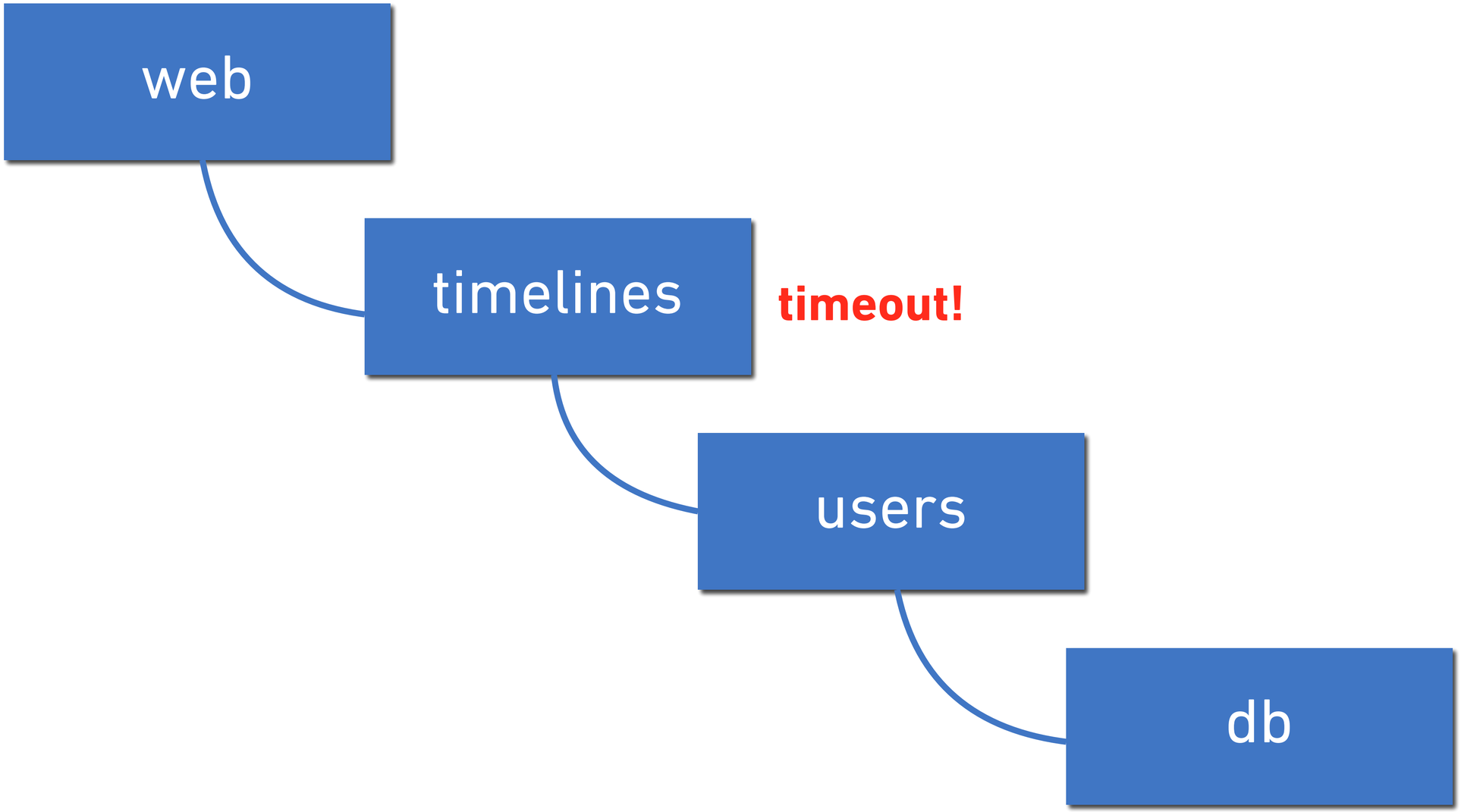
worst-case: 300% more load!!!

better:

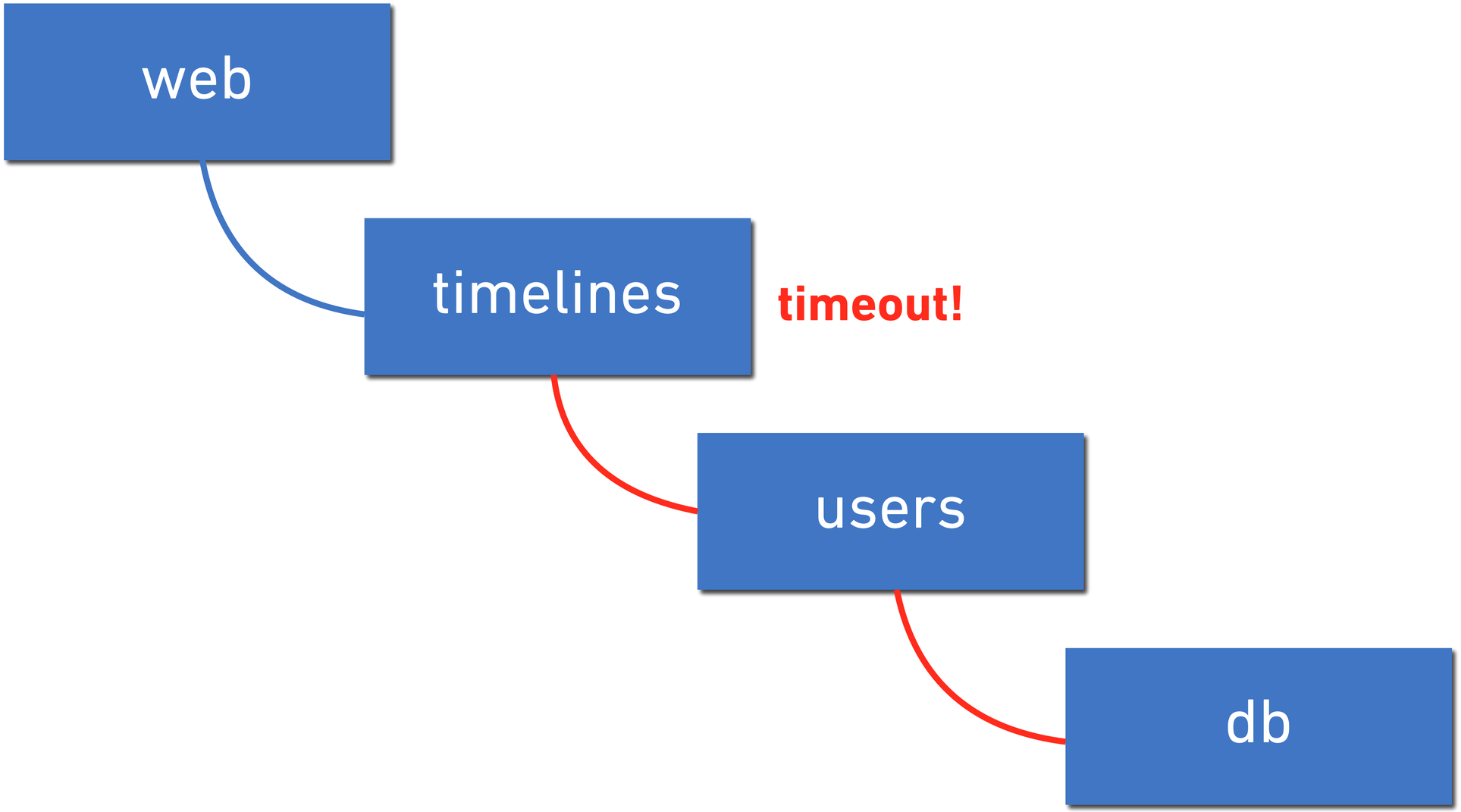
retryBudget=20%

worst-case: 20% more load

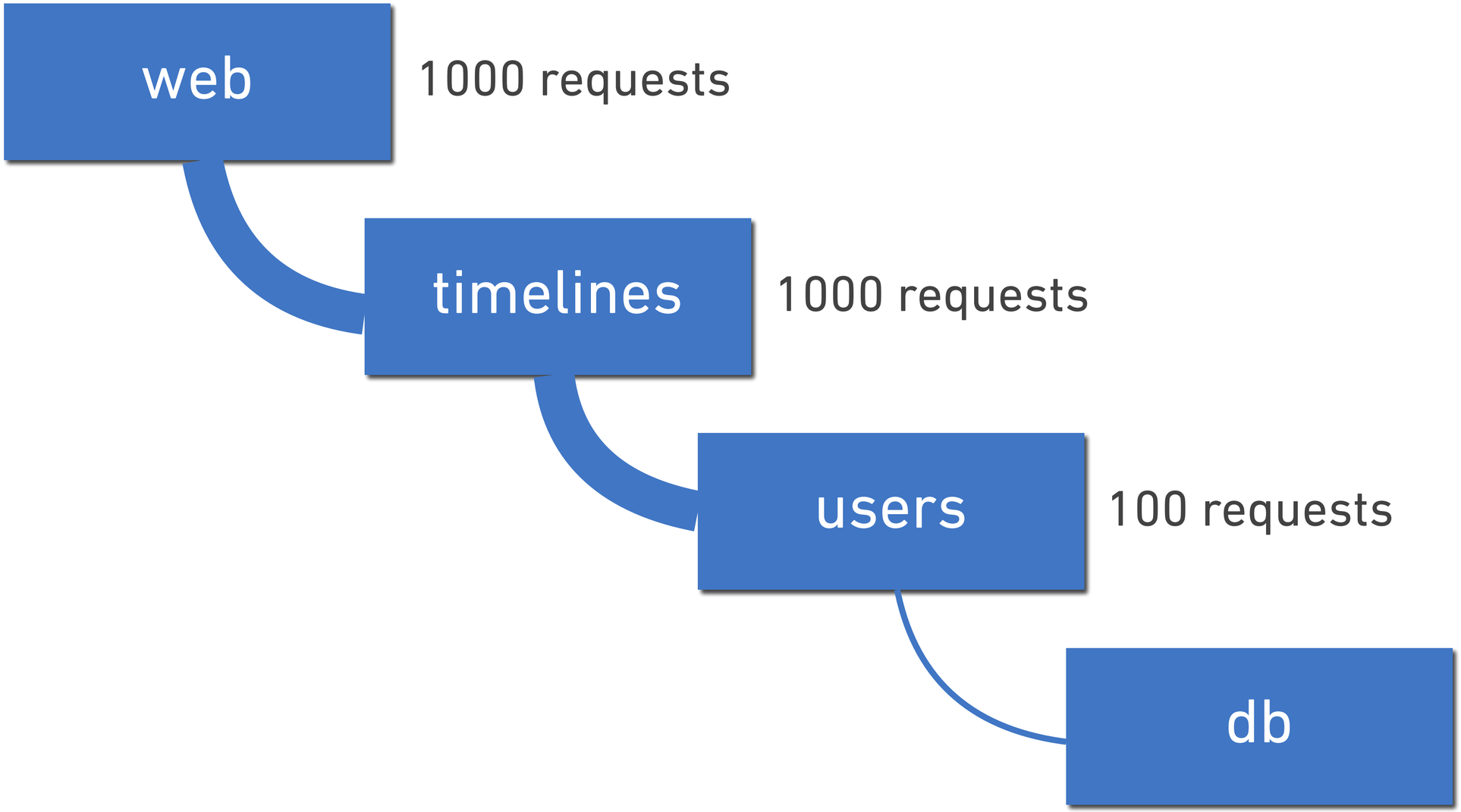
load shedding via cancellation



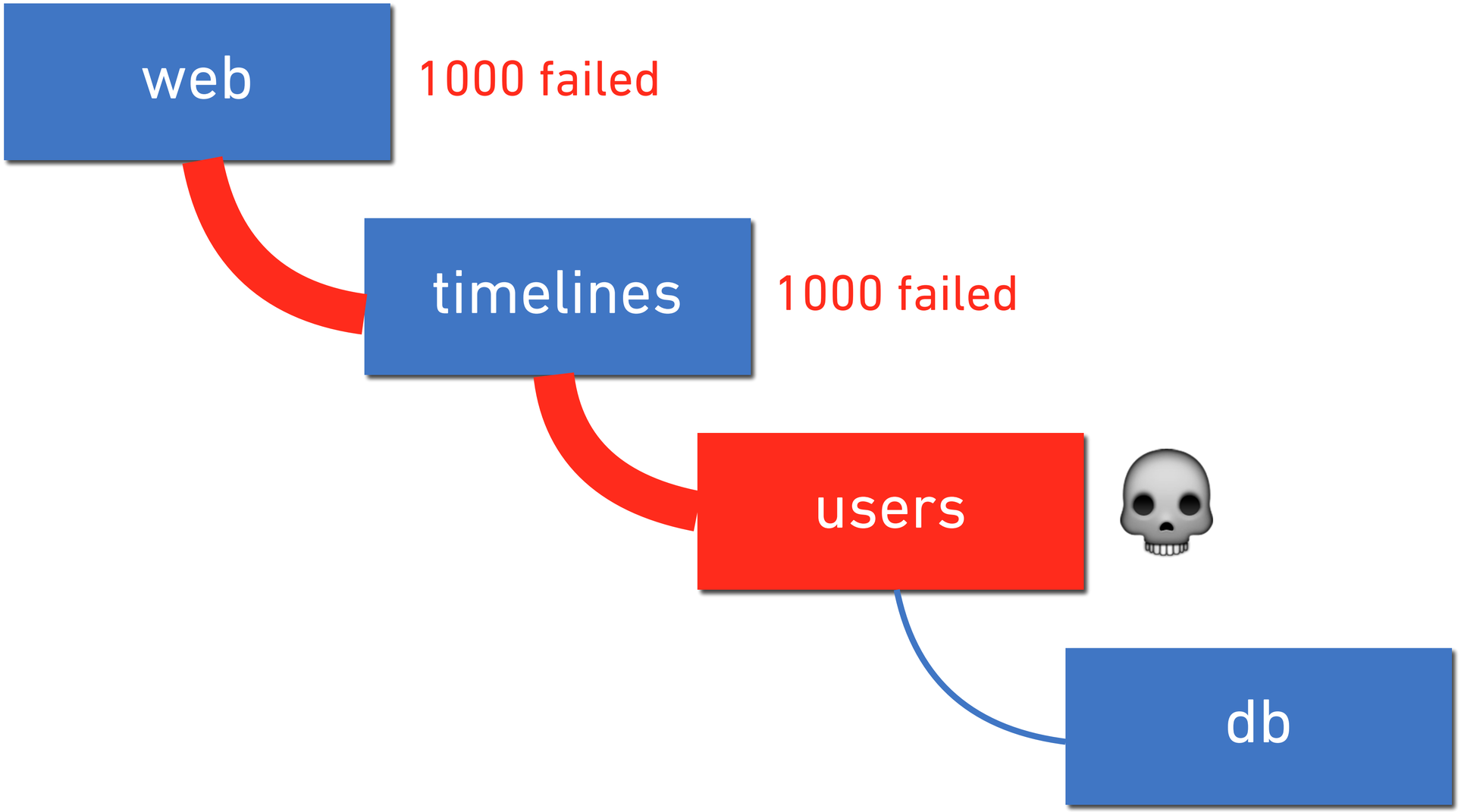
load shedding via cancellation



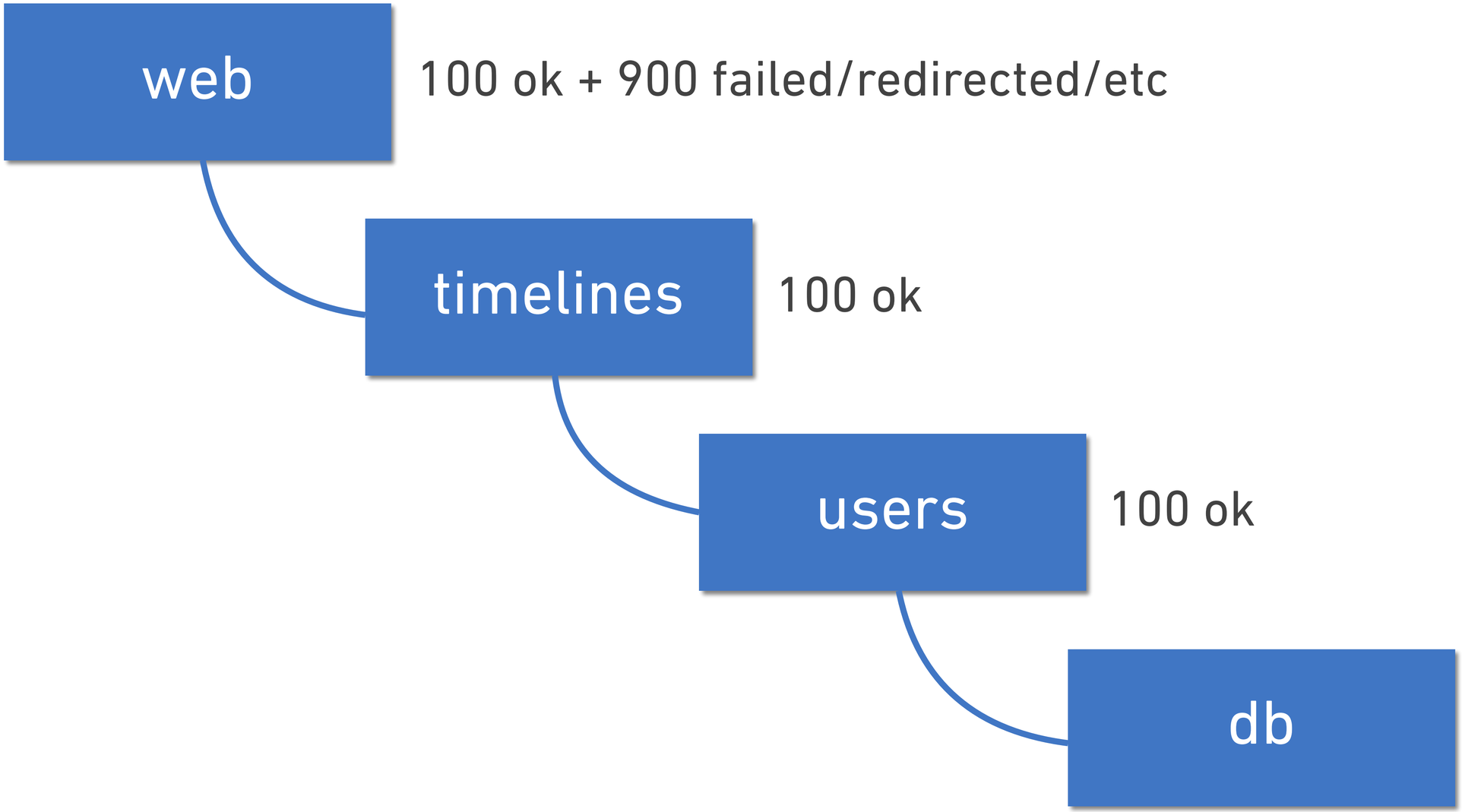
backpressure



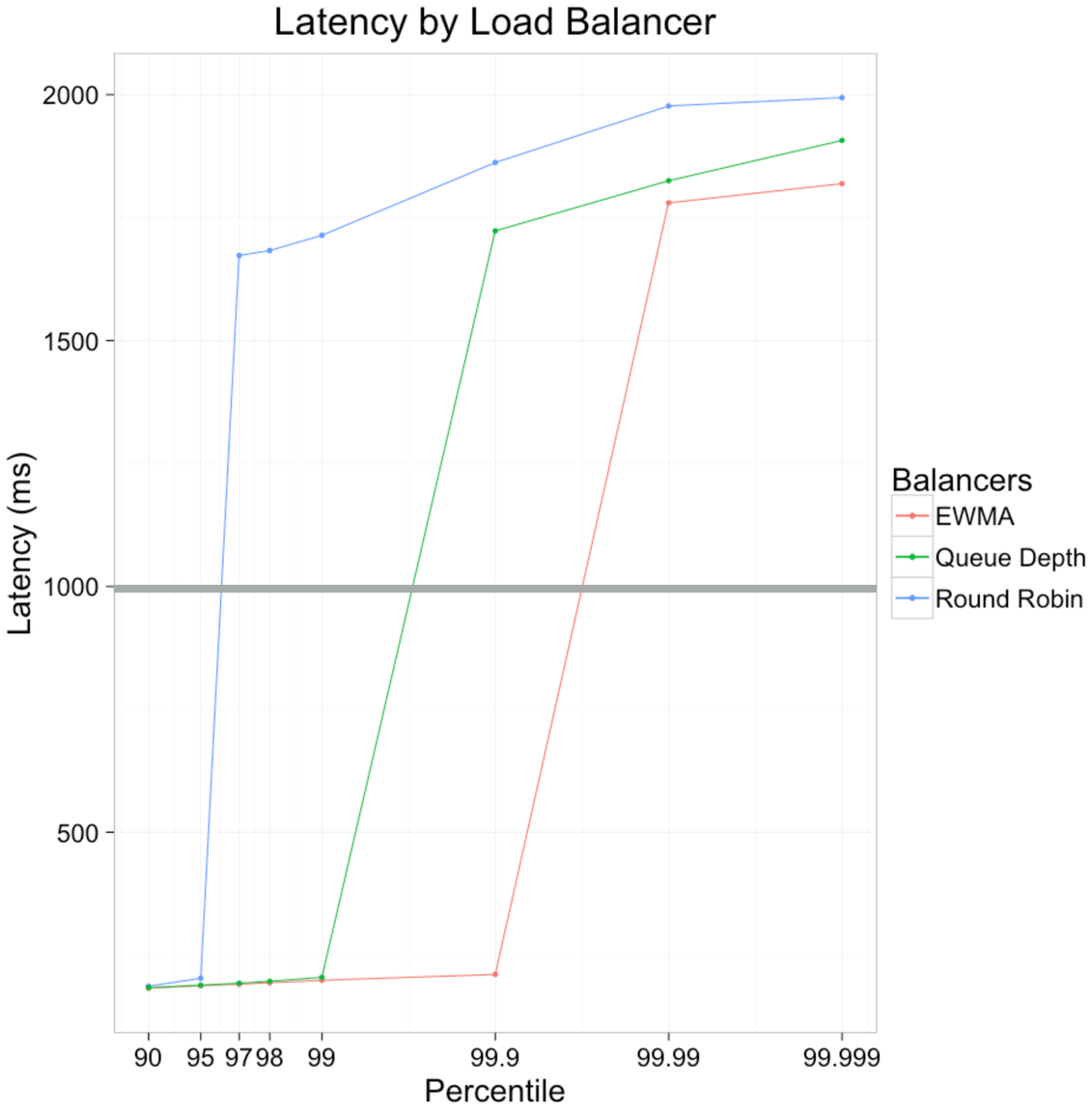
backpressure



backpressure



request-level load balancing



lb algorithms:

- ~~round-robin~~
- ~~fewest connections~~
- queue depth
- exponentially-weighted moving average (ewma)
- aperture

Phil Calçado
@pcalcado

Follow

Any sufficiently complicated microservices arch contains an ad-hoc, informally-specified, bug-ridden, slow implementation of half of Finagle

4:19 AM - 29 Jun 2016 · Manhattan, NY, United States

31 63

So just rewrite everything in Finagle!?

linkerd

github.com/buoyantio/linkerd

service mesh proxy

built on **finagle & netty**

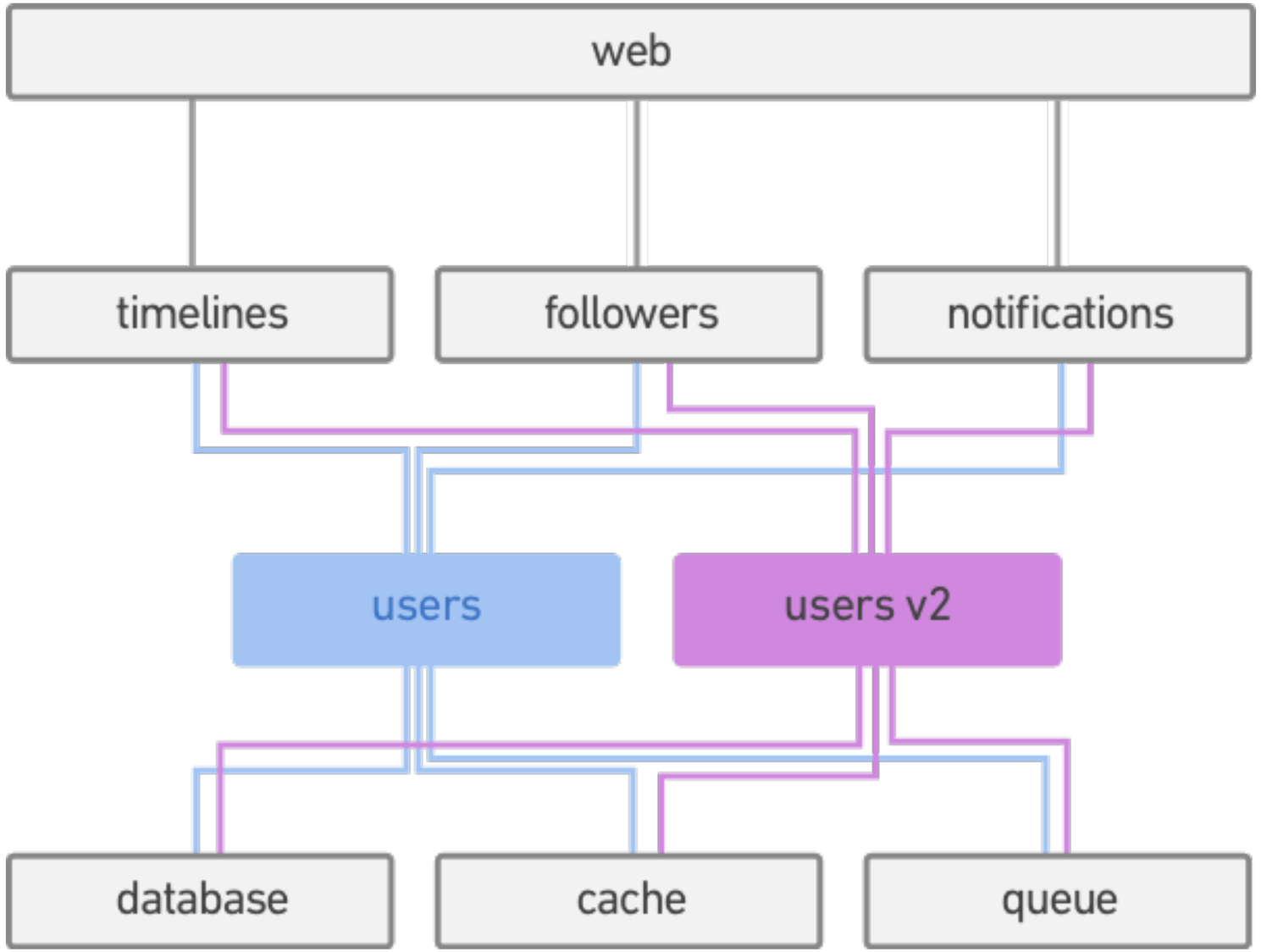
suuuuper pluggable

- http, thrift, ...
- etcd, consul, kubernetes, marathon, zookeeper, ...
- ...

1.1 What Do Linkers and Loaders Do?

The basic job of any linker or loader is simple: It binds more abstract names to more concrete names, which permits programmers to write code using the more abstract names. For example, it takes a name written by a programmer such as `getline` and binds it to "the location 612 bytes from the beginning of the executable code in module `iosys`." Or it may take a more abstract numeric address such as "the location 450 bytes beyond the beginning of the static data for this module" and bind it to a numeric address.

linker for the datacenter



logical naming

applications refer to *logical names*

requests are bound to *concrete names*

delegations express routing

```
/s/users
```

```
/#/io.l5d.zk/prod/users
/#/io.l5d.zk/staging/users
```

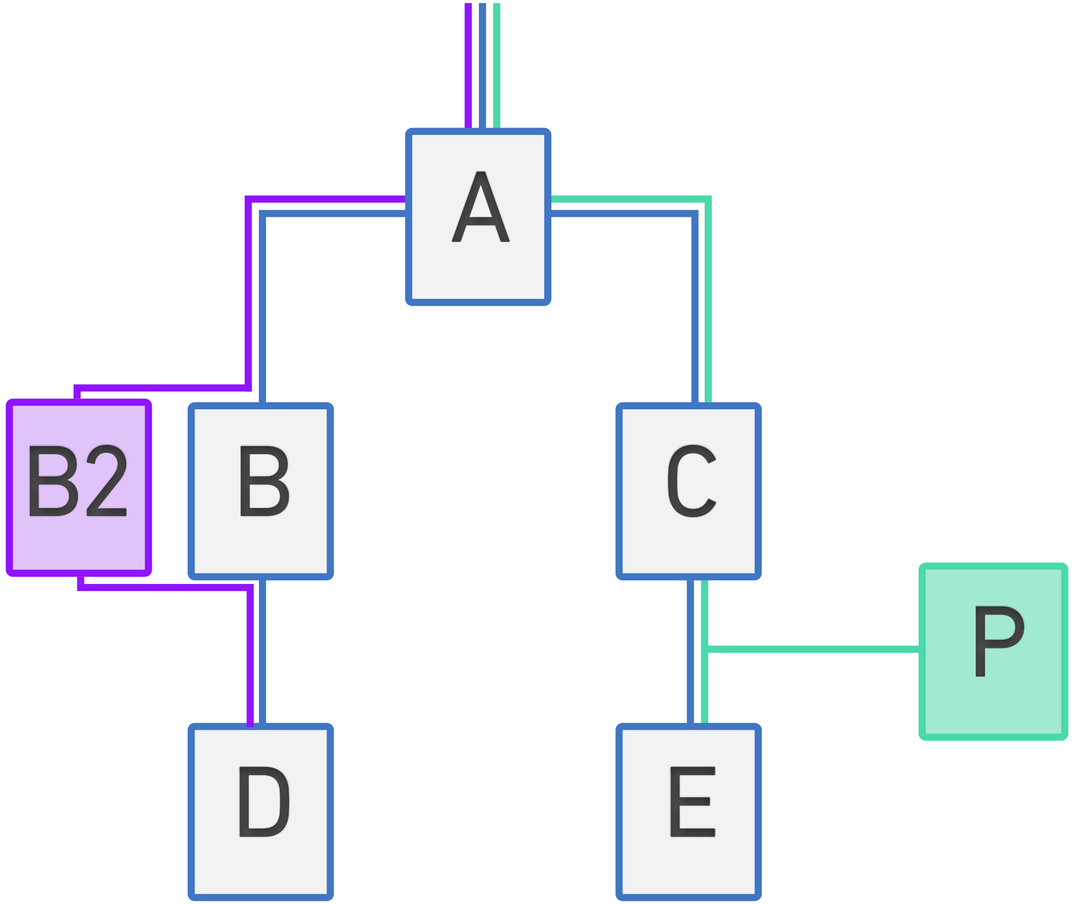
```
/s => /#/io.l5d.zk/prod
```

per-request routing: staging

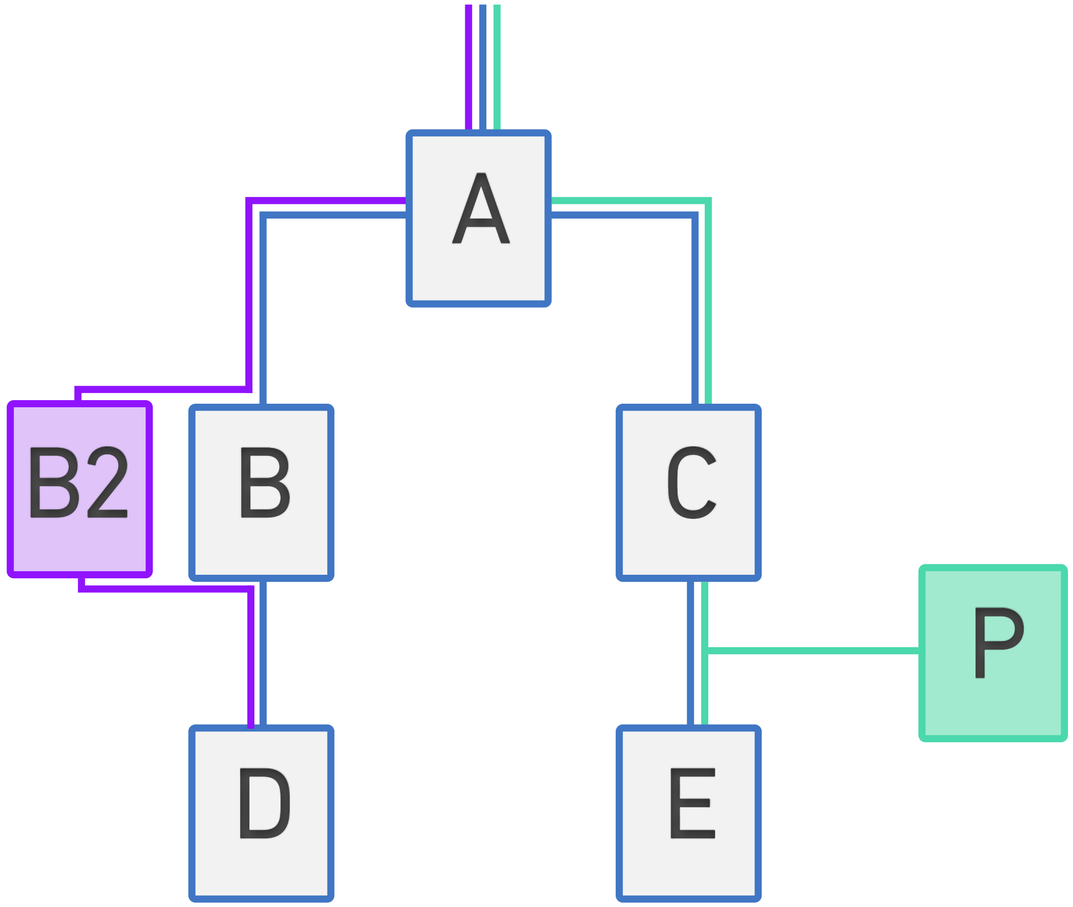
GET / HTTP/1.1

Host: mysite.com

l5d-dtab: /s/B => /s/B2



per-request routing: debug proxy



GET / HTTP/1.1

Host: mysite.com

l5d-dtab: /s/E => /s/P/s/E

linkerd service mesh

transport security

service discovery

circuit breaking

backpressure

deadlines

retries

tracing

metrics

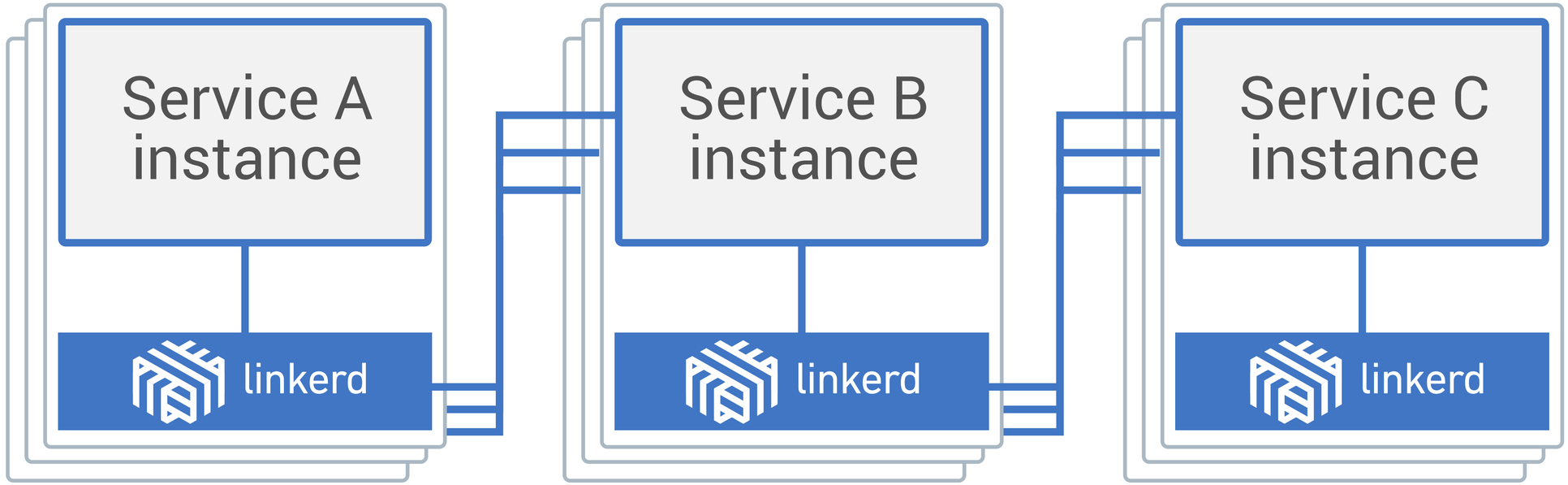
keep-alive

multiplexing

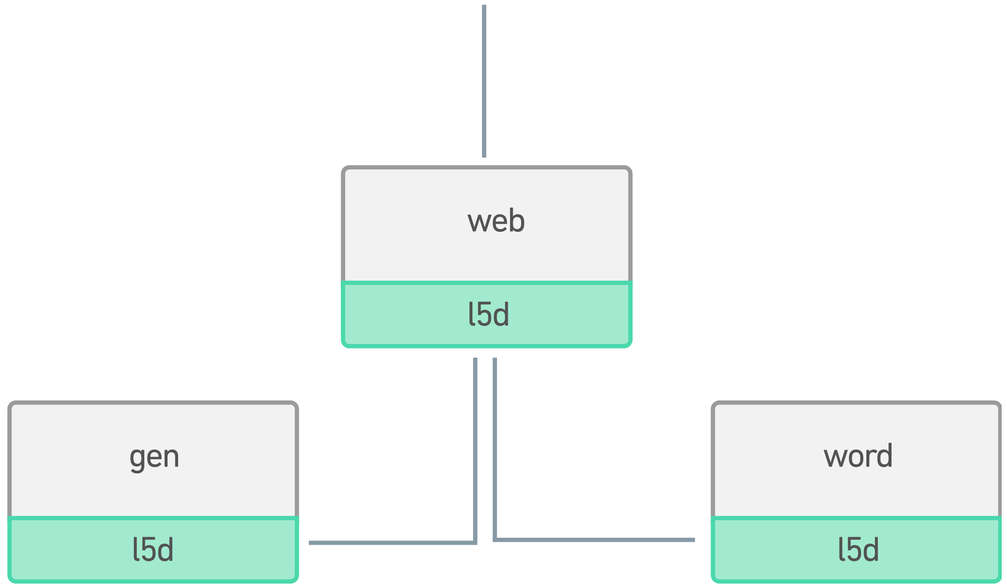
load balancing

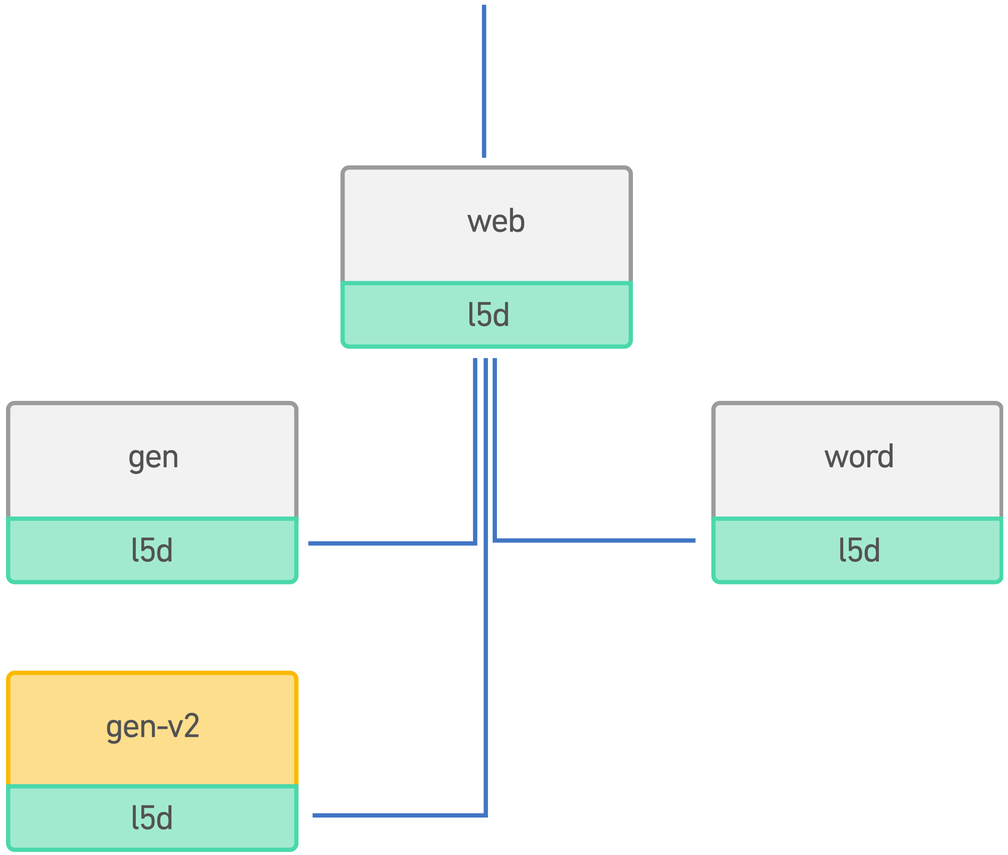
per-request routing

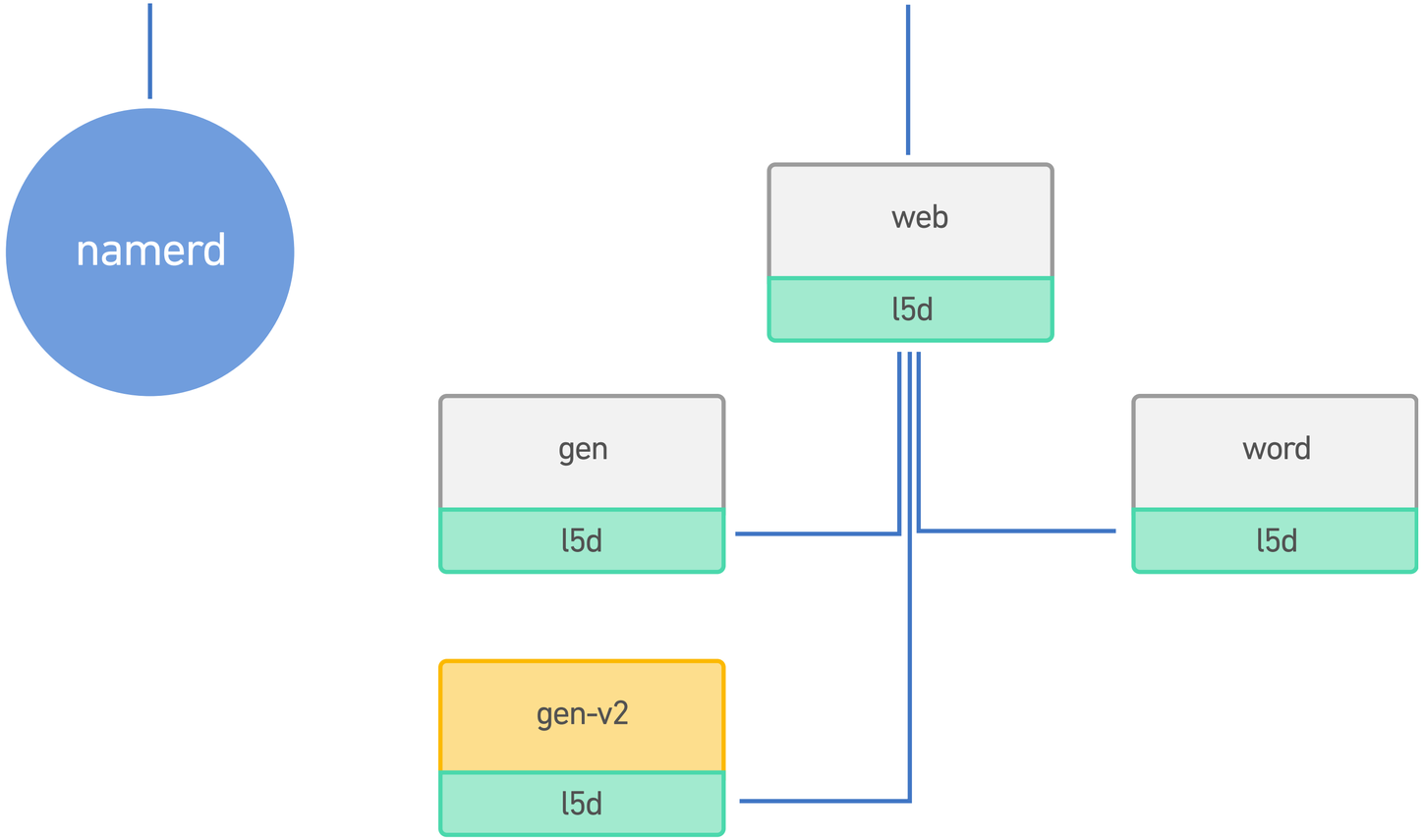
service-level objectives



demo: gob's microservice


web
l5d
gen
l5d
word
l5d

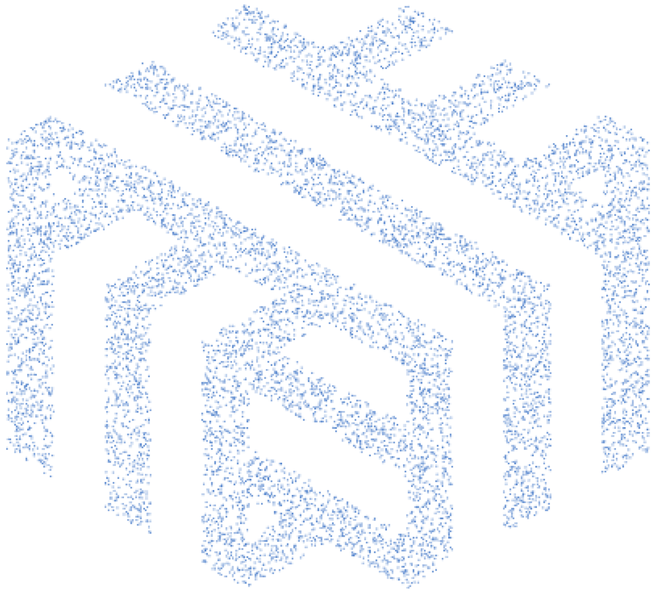

web
l5d
gen
l5d
word
l5d
gen-v2
l5d


namerd
web
l5d
gen
l5d
word
l5d
gen-v2
l5d

github.com/buoyantio/linkerd-examples

linkerd roadmap

- Battle test HTTP/2
- TLS client certs
- Deadlines
- Dark Traffic
- *All configurable everything*



thanks!

more at **linkerd.io**

slack: **slack.linkerd.io**

email: ver@buoyant.io

twitter:

- @olix0r
- @linkerd