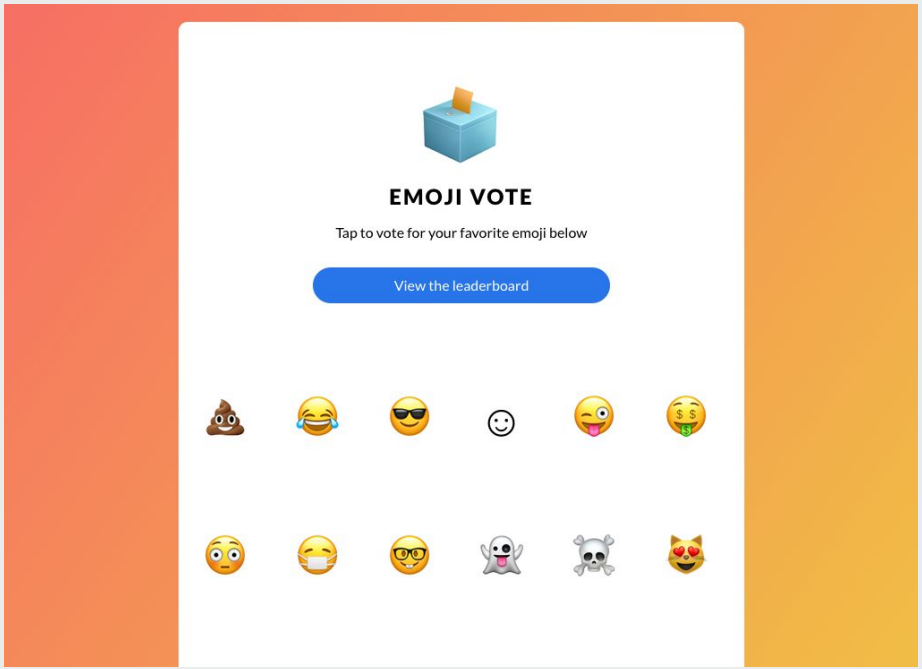
Scale Your Service on What Matters: Autoscaling on Latency

Get your votes in

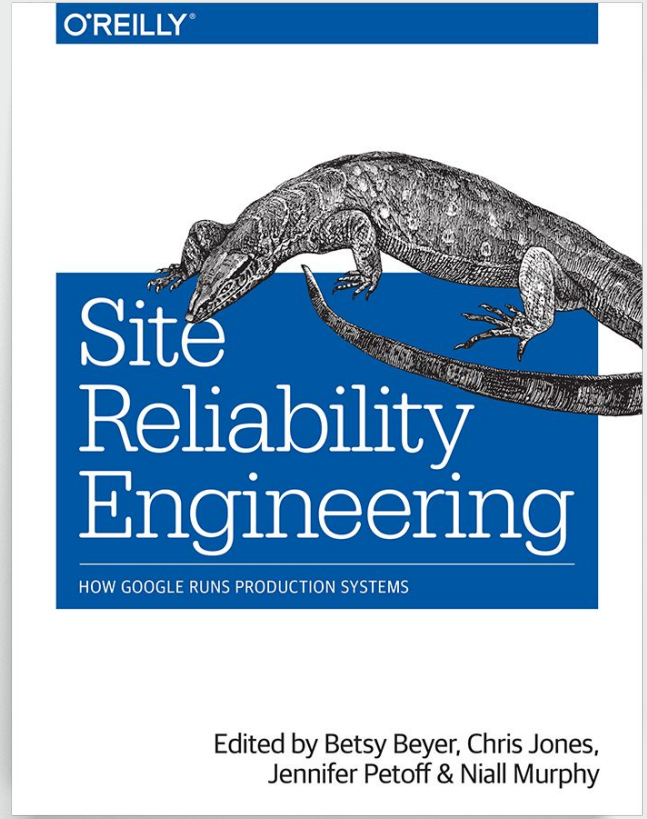


https://kc.l5d.io

TODO: SOMETHING

Golden Signals

- Latency
- Traffic
- Errors
- Saturation



CPU is an approximation

Environments are complex

Every request matters

- Tail latency is important
-

Site	# of requests	page loads that would experience the 99%'lie [(1 - (.99 ^ N)) * 100%]
amazon.com	190	85.2%
kohls.com	204	87.1%
jcrew.com	112	67.6%
saksfifthavenue.com	109	66.5%
--	--	--
nytimes.com	173	82.4%
cnn.com	279	93.9%
--	--	--
twitter.com	87	58.3%
pinterest.com	84	57.0%
facebook.com	178	83.3%
--	--	--
google.com (yes, that simple noise-free page)	31	26.7%
google.com search for "http requests per page"	76	53.4%

TODO: SOMETHING

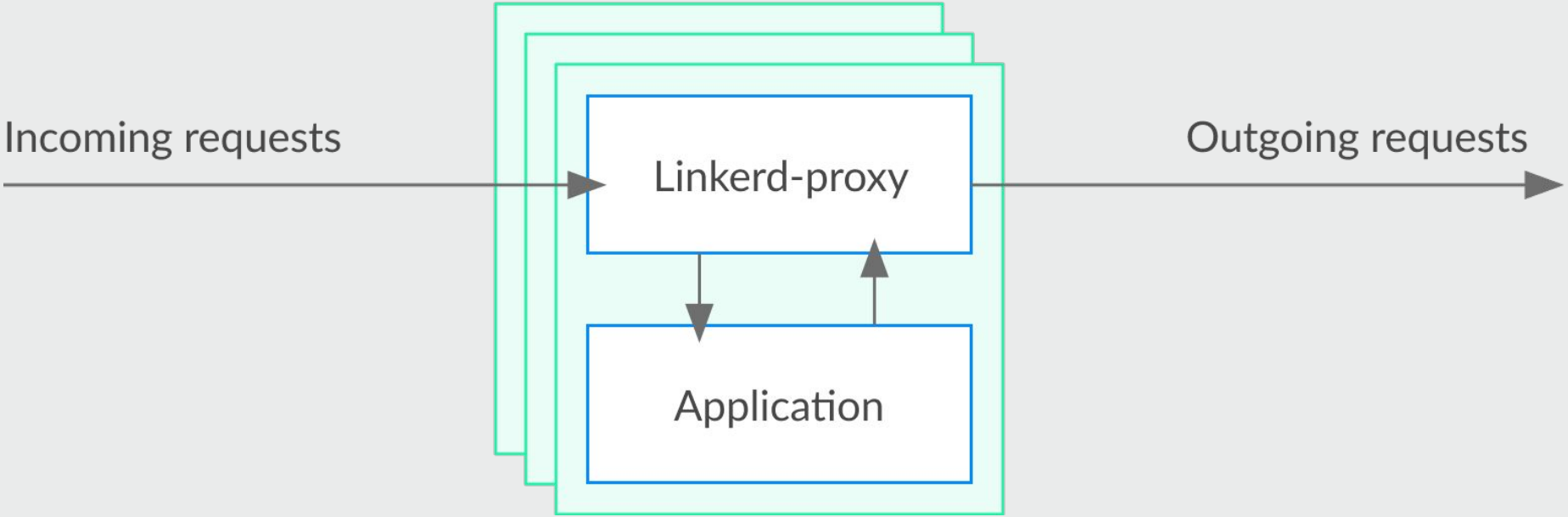
What is required?

- [ ] Measure the latency of a service
- [ ] Expose custom metrics
- [ ] Autoscale!

An open source *service mesh* and CNCF member project.

- 24+ months in production
- 2,500+ Slack members
- 7,500+ GitHub stars
- 40m+ DockerHub pulls
- 100+ contributors
- 400b+ requests/mo

What is Linkerd?



Architecture

FIXME

DEMO: Install Linkerd

What is required?

- [x] Measure the latency of a service
- [ ] Expose custom metrics
- [ ] Autoscale!

What are custom metrics?

Architecture

FIXME

DEMO: Expose custom metrics

What is required?

- [x] Measure the latency of a service
- [x] Expose custom metrics
- [ ] Autoscale!

Architecture

FIXME

DEMO: Autoscaling

What is required?

- ✅ Measure the latency of a service
- ✅ Expose custom metrics
- ✅ Autoscale!

Route Based Scaling

Predictive Scaling

Slides	http://bit.ly/linkerd-tidesf
Tutorial	http://bit.ly/tidesf-tutorial
Get Started!	https://bit.ly/linkerd-get-started
Kubecon	http://bit.ly/linkerd-kubecon