Project Frankenstein

A multi-tenant, horizontally scalable Prometheus as a Service

Tom Wilkie (& Julius Volz)
Weaveworks, August 2016

FRANKENSTEIN;

OR,

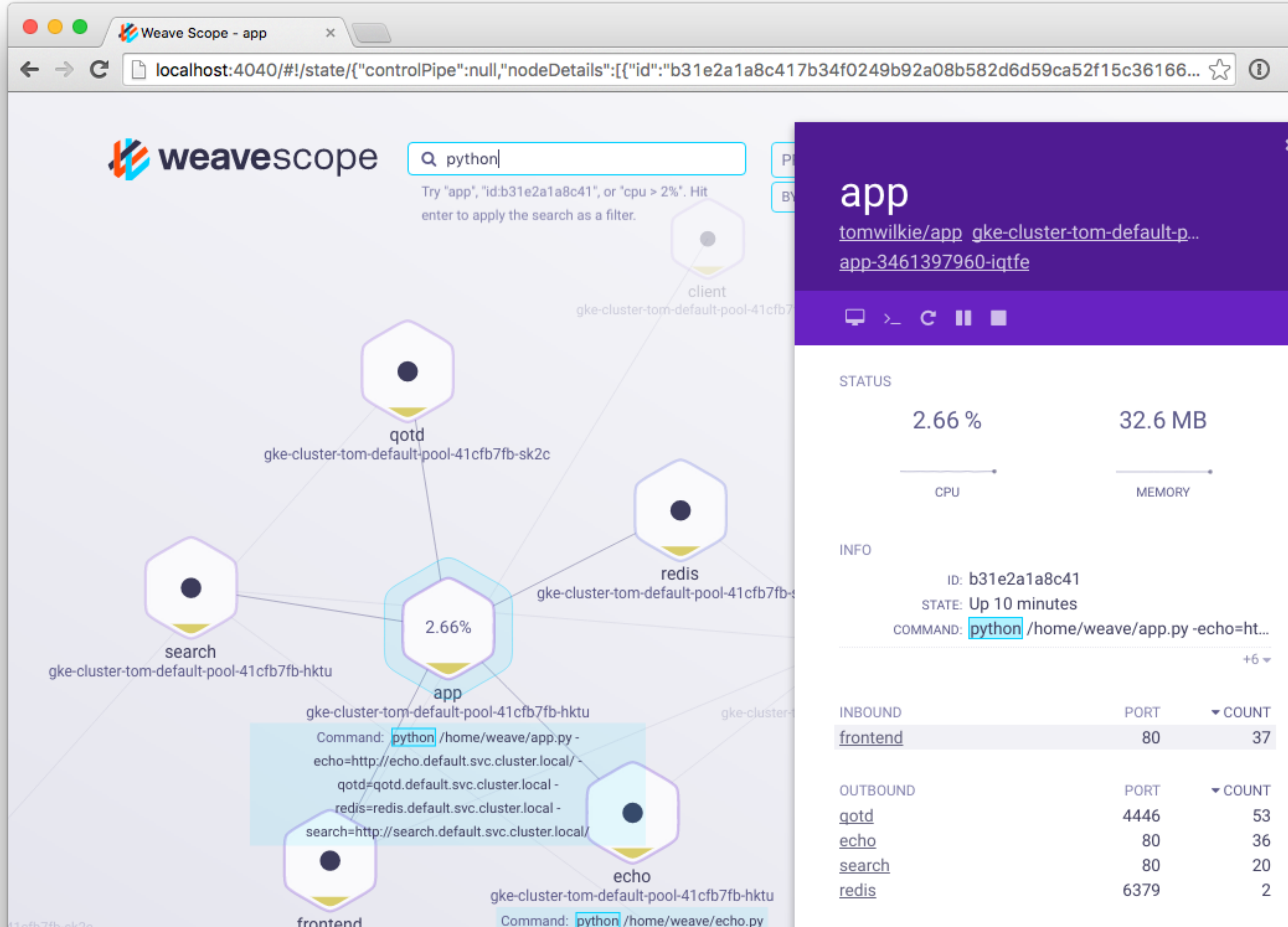
THE MODERN PROMETHEUS.

BY MARY W. SHELLY,

AUTHOR OF 'THE LAST MAN,' 'PERKIN WARBECK,' &C.

weaveworks

“the best way to visualise, manage & monitor your cloud native application”


Weave Scope - app
localhost:4040/#!/state/{"controlPipe":null,"nodeDetails":[{"id":"b31e2a1a8c417b34f0249b92a08b582d6d59ca52f15c36166...
weavescope
python
Try "app", "id:b31e2a1a8c41", or "cpu > 2%". Hit enter to apply the search as a filter.
client
qotd
gke-cluster-tom-default-pool-41cfb7fb-sk2c
redis
search
gke-cluster-tom-default-pool-41cfb7fb-hktu
2.66%
app
gke-cluster-tom-default-pool-41cfb7fb-hktu
echo
gke-cluster-tom-default-pool-41cfb7fb-hktu
frontend

Command: python /home/weave/app.py -echo=http://echo.default.svc.cluster.local/ -qotd=qotd.default.svc.cluster.local -redis=redis.default.svc.cluster.local -search=http://search.default.svc.cluster.local/

Command: python /home/weave/echo.py

app

tomwilkie/app gke-cluster-tom-default-p...

app-3461397960-iqtfe

STATUS

2.66 % CPU

32.6 MB MEMORY

INFO

ID: b31e2a1a8c41

STATE: Up 10 minutes

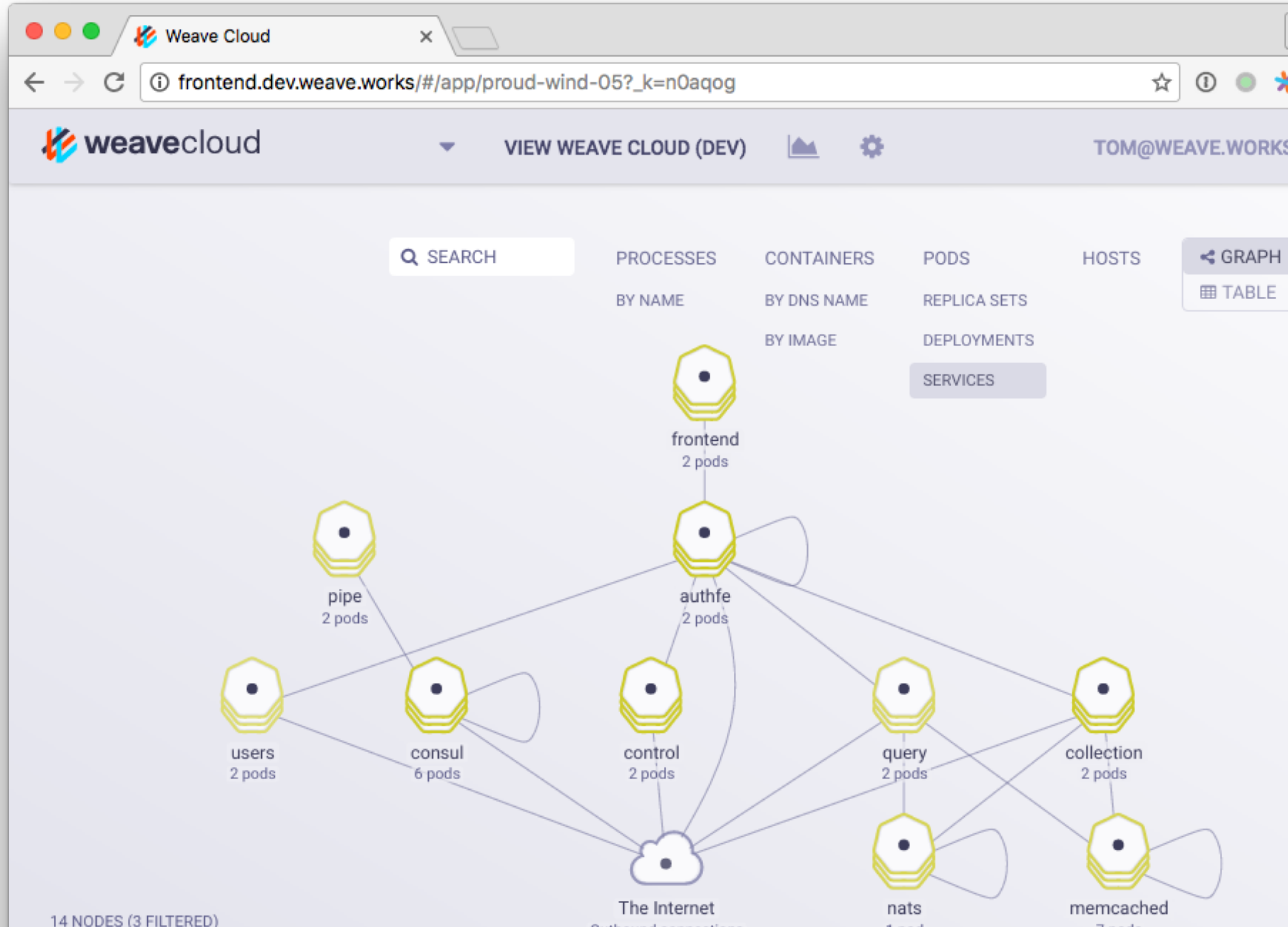
COMMAND: python /home/weave/app.py -echo=ht...

+6

INBOUND	PORT	COUNT
frontend	80	37

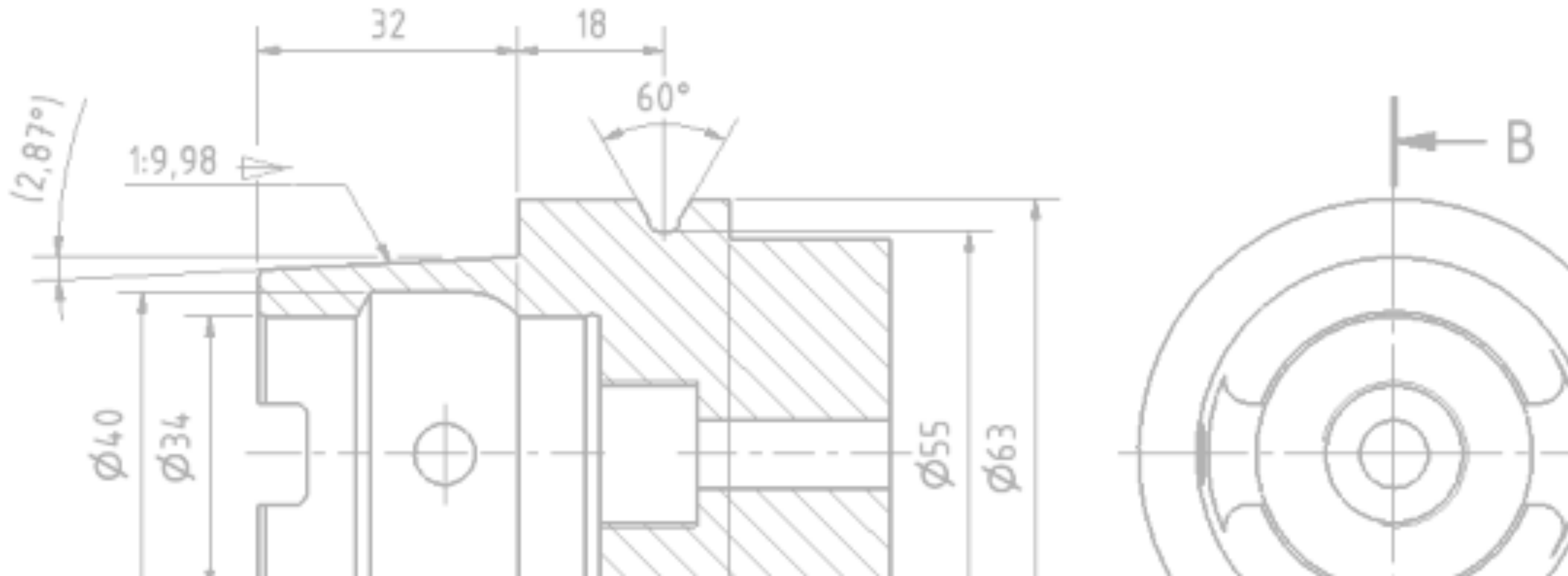
OUTBOUND	PORT	COUNT
qotd	4446	53
echo	80	36
search	80	20
redis	6379	2

weaveworks


Weave Cloud
frontend.dev.weave.works/#/app/proud-wind-05?_k=n0aqog
weavecloud
VIEW WEAVE CLOUD (DEV)
TOM@WEAVE.WORKS
SEARCH
PROCESSES
BY NAME
CONTAINERS
BY DNS NAME
BY IMAGE
PODS
REPLICA SETS
DEPLOYMENTS
SERVICES
HOSTS
GRAPH
TABLE
frontend
2 pods
pipe
2 pods
authfe
2 pods
users
2 pods
consul
6 pods
control
2 pods
query
2 pods
collection
2 pods
The Internet
nats
memcached
14 NODES (3 FILTERED)

weaveworks

Design



weaveworks

why not just run my own Prometheus?

- the as-a-service bit provides authentication and access control
- virtually infinite retention; all the state is managed for you, by us
- provide a different story around durability, HA and scalability
- (eventually) better query performance, especially for long queries

weaveworks

requirements:

1. API compatible with Prometheus
2. easy to operate and manage
3. tens of thousands of users, tens of millions samples/s
4. cost effective to run
5. reuse as much of Prometheus as possible

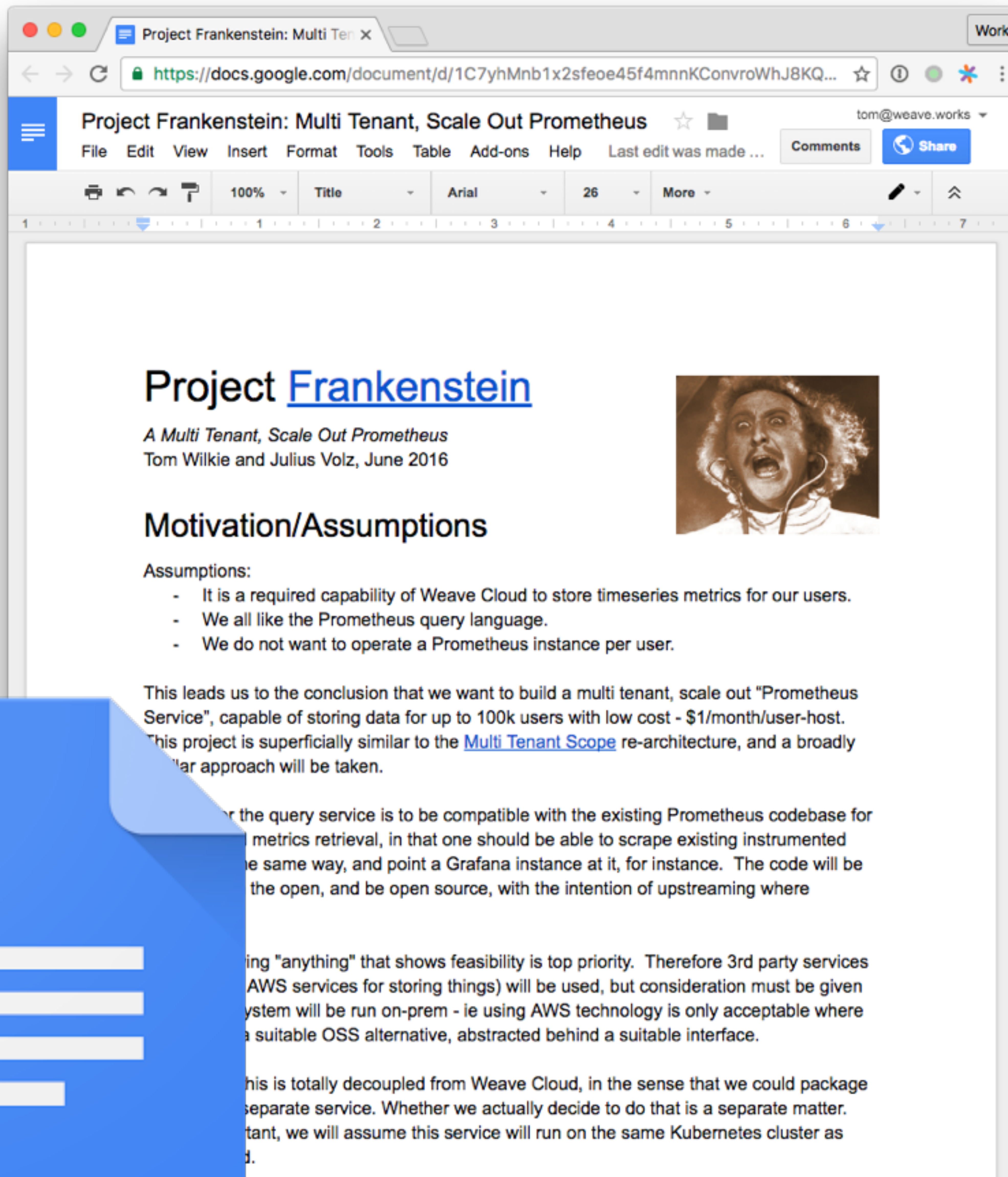
… so we can sell it

weaveworks

Aim: build proof of concept as quickly as possible

http://goo.gl/prdUYV

16/06	started design doc
22/06	circulated on list
22/06	initial commit
26/07	launch jobs
25/08	give talk!


Project Frankenstein: Multi Te
https://docs.google.com/document/d/1C7yhMnb1x2sfeoe45f4mnnKConvroWhJ8KQ...
Project Frankenstein: Multi Tenant, Scale Out Prometheus
tom@weave.works
File Edit View Insert Format Tools Table Add-ons Help Last edit was made ...
Comments
Share
100% Title Arial 26 More

Project Frankenstein

A Multi Tenant, Scale Out Prometheus
Tom Wilkie and Julius Volz, June 2016

Motivation/Assumptions

Assumptions:
- It is a required capability of Weave Cloud to store timeseries metrics for our users.
- We all like the Prometheus query language.
- We do not want to operate a Prometheus instance per user.

This leads us to the conclusion that we want to build a multi tenant, scale out "Prometheus Service", capable of storing data for up to 100k users with low cost - $1/month/user-host. This project is superficially similar to the Multi Tenant Scope re-architecture, and a broadly ...lar approach will be taken.

...r the query service is to be compatible with the existing Prometheus codebase for ... metrics retrieval, in that one should be able to scrape existing instrumented ...e same way, and point a Grafana instance at it, for instance. The code will be ... the open, and be open source, with the intention of upstreaming where

...ing "anything" that shows feasibility is top priority. Therefore 3rd party services ...AWS services for storing things) will be used, but consideration must be given ...ystem will be run on-prem - ie using AWS technology is only acceptable where ... suitable OSS alternative, abstracted behind a suitable interface.

...his is totally decoupled from Weave Cloud, in the sense that we could package ...eparate service. Whether we actually decide to do that is a separate matter. ...tant, we will assume this service will run on the same Kubernetes cluster as ...d.

weaveworks

Your DC
Retriever
scraping
your jobs
Weave Cloud
Frontend, Authenticator
Distributor
…
Distributor
Ingester
Ingester
Ingester
DynamoDB
S3

Retriever

Does scraping and relabelling.

Is a vanilla Prometheus plus:

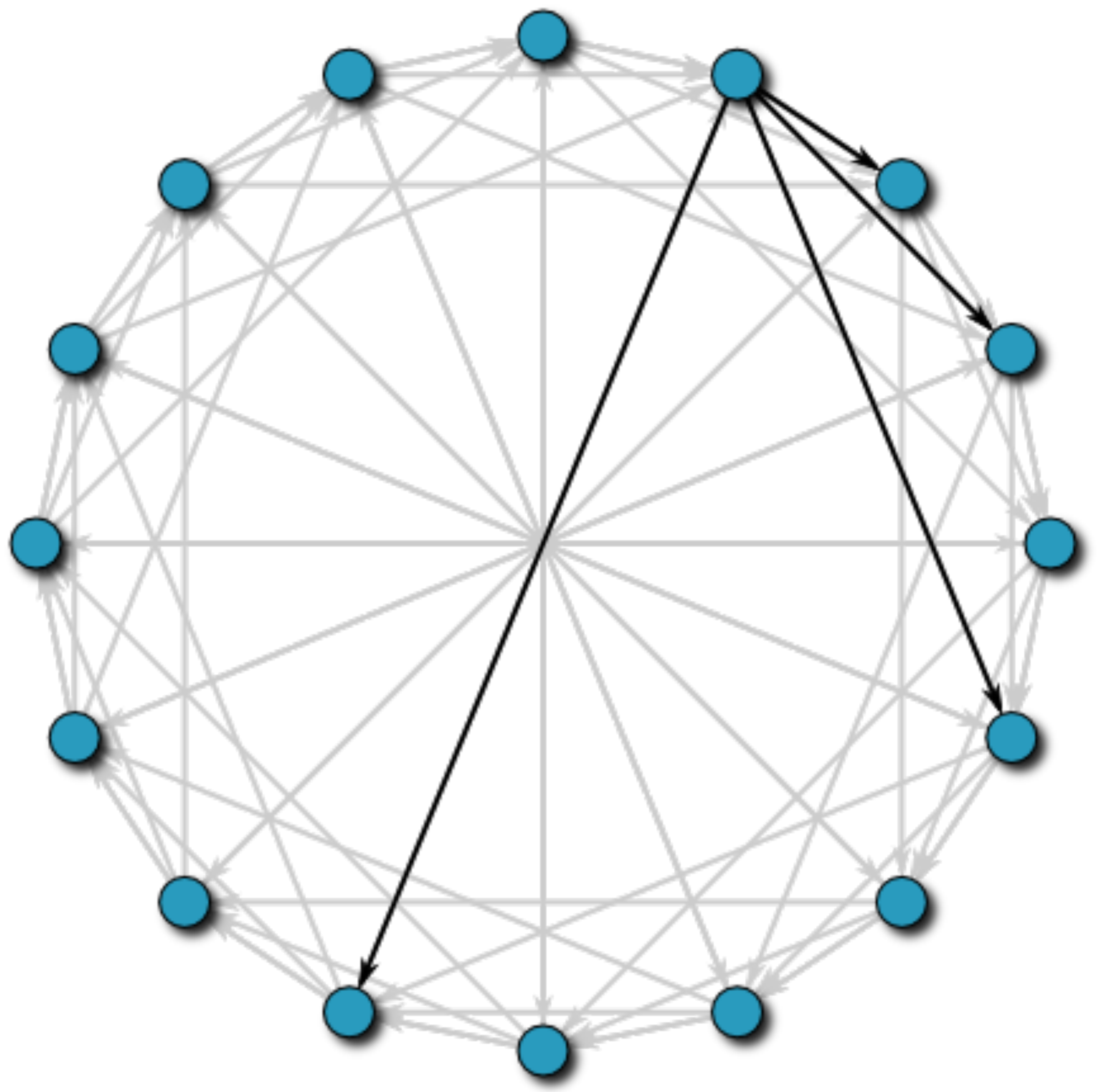
- Brian Brazil's generic write PR (#1487)
- Some modification to prevent local storage + indexing

```
/bin/prometheus -retrieval-only -storage.remote.generic-url=...
```

weaveworks

Distributor

- Uses consistent hashing to assign timeseries to Ingesters
- Input to hash is (user ID, metric name)
- Tokens stored in Consul
- Also currently handles queries

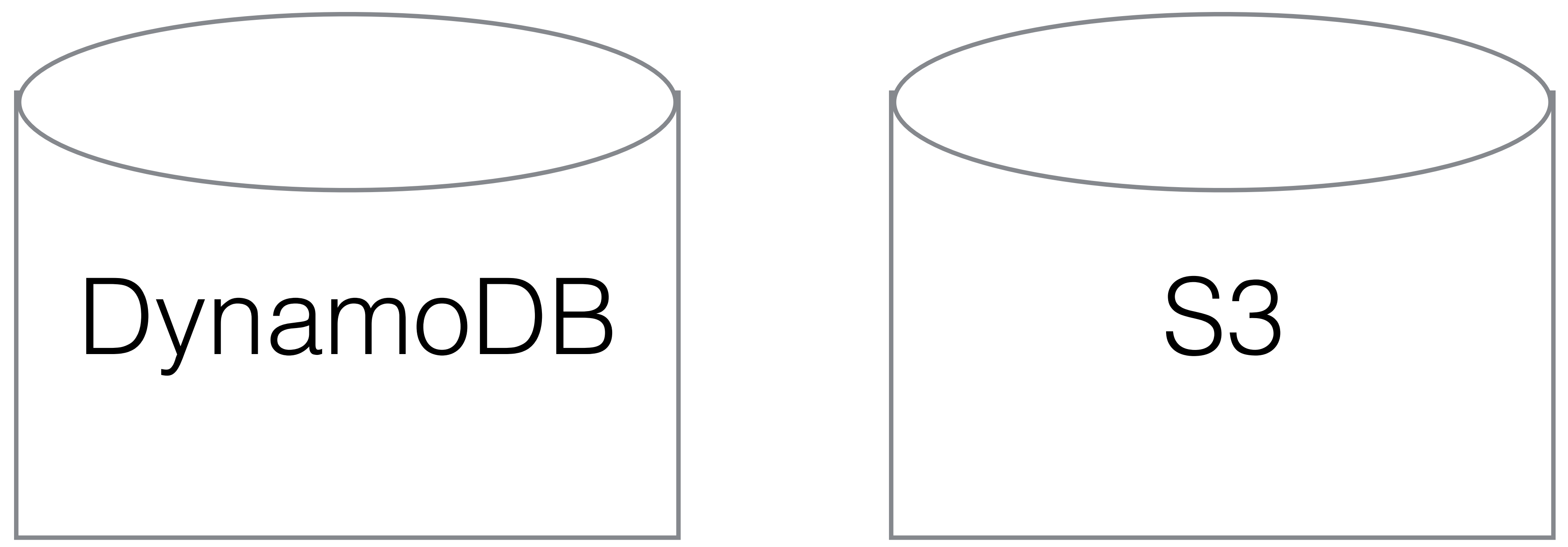


http://goo.gl/U9u1U2

weaveworks

Ingester

- Heavily modified MemorySeriesStorage
- Use same chunk format as Prometheus
- Keeps everything in memory (for up to an hour)
- Also stores in memory inverted index for queries
- Flushes chunks to S3 and indexes them in DynamoDB



External inverted index maintained in DynamoDB, chunks stored in S3

Item in DynamoDB looks like:

```
{
  hash key: "{user ID}:{metric name}:{hour}",
  range key: "{label name}:{label value}:{chunk ID}",
  metric: ...,
  from, through: ...,
  ID: ...,
}
```

weaveworks

Evaluation

The Good

- It works! And in ~2 months.
- Seems pretty scalable, handling two clusters right now
- Query performance better than expected

The Bad

- Hashing scheme means can’t do queries that don’t involve metric names.
- Possible to hotspot an ingester

The Ugly: the code…

weaveworks

Demo

weaveworks

Lots left to do…

Features:

- Recording rules
- Alerting & Alertmanager

Reliability:

- Replication between ingesters, commit log etc
- Ingestor lifecycle
- Separate query service?

Performance:

- Query parallelisation
- Background chunk coalescing

Code:

- Code cleanup
- Upstream appropriate changes

weaveworks

Questions?

https://github.com/tomwilkie/prometheus

Try it out!

Email help@weave.works for
instructions and to get on white list

weaveworks