AUTOSCALING A SIMPLE WEB SERVER WITH KUBERNETES

OBLIGATORY ABOUT ME

> SENIOR SOFTWARE ENGINEER AT MAVENLINK
> BOOTCAMP GRAD
> KOOB NOOB
> RECENT KUBECON ATTENDEE

A TRUEISH STORY...

BUZZ

> A LITTLE ANXIOUS
> LIKES LONG WALKS ON THE BEACH
> LIKES DESTROYING TOYS

I WANT TO FIND FRIENDS FOR BUZZ

TINDER FOR DOGS*

A LITTLE BIT OF TRAFFIC...

HOSTED ON HEROKU...

ONE DEVELOPER...

WHEN SUDDENLY...

6.4k

Check this out! Tinder for Dogs https://tinderfordogs.com... ⤤

Posted by u/ashthecat15 🔖5 hours ago

⤢ | 💬 **43 Comments** ➤ **Share** •••

sine.com (DNS queries per minute)

website first linked to on Reddit

0

10/13 10:00   10/13 12:00   10/13 14:00   10/13 16:00   10/13 18:00   10/13 20:00   10/13 22:00

503 Service Temporarily Unavailable

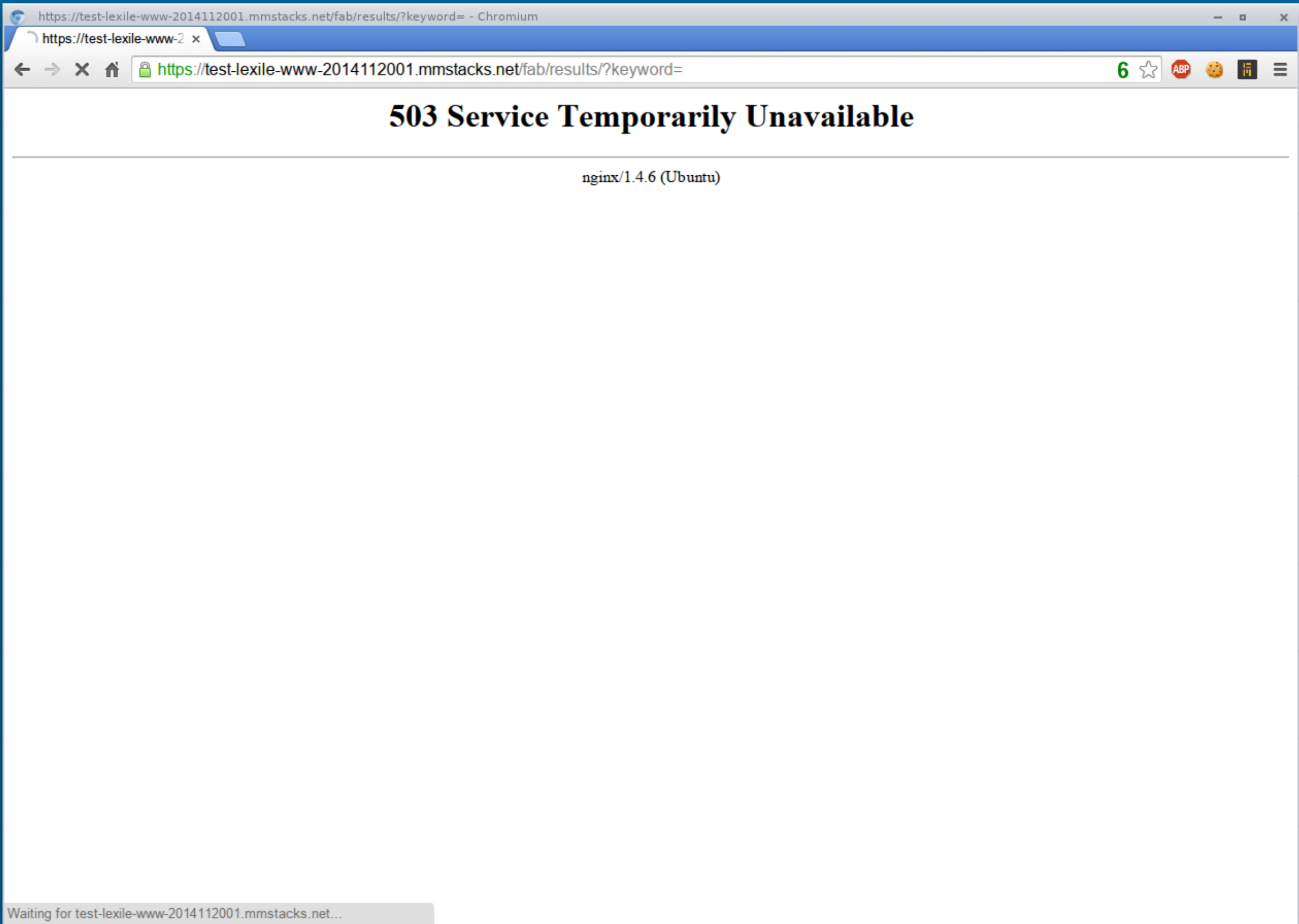nginx/1.4.6 (Ubuntu)
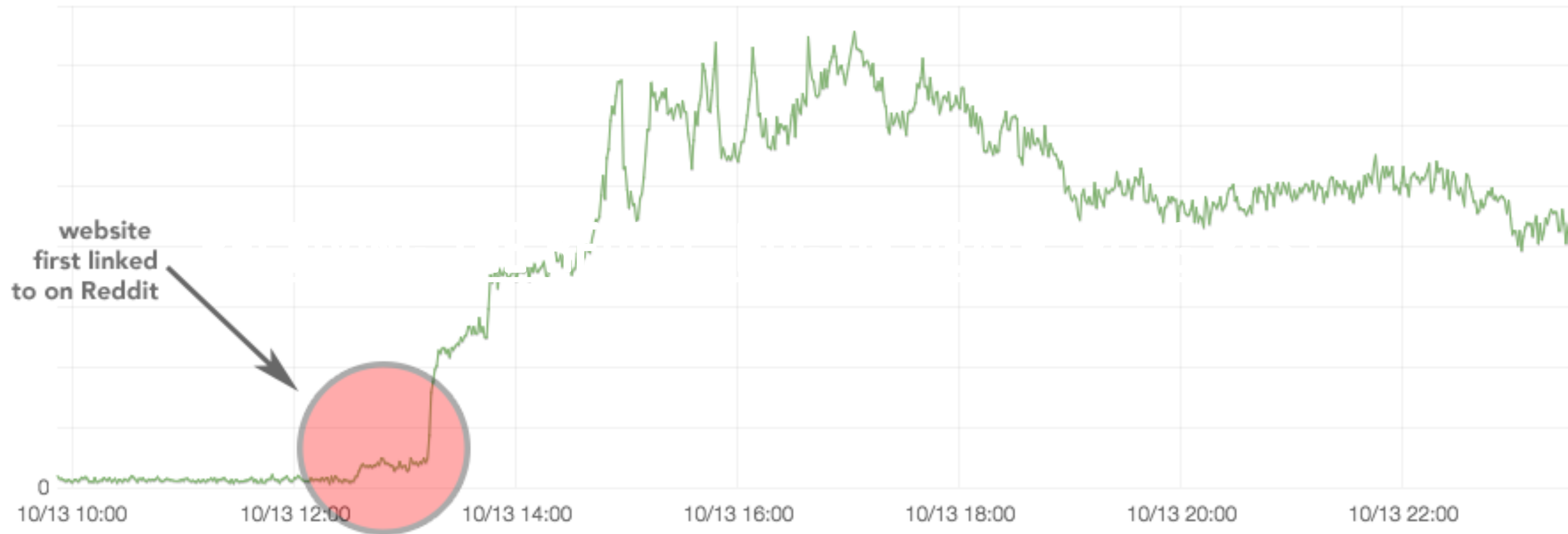
OUR DEVELOPER IS SCRAMBLING TO KEEP THE SITE ONLINE...

INVESTORS ARE CALLING...

OUR CEO IS FURIOUS THAT WE CAN'T SERVE ALL THESE NEW CUSTOMERS...

... AND DOGS CAN'T FIND THE LOVE THEY DESERVE.

HOW CAN WE STAY ONLINE DURING INCREASED LOAD?

IDEALLY WITH AN OPEN SOURCE FRAMEWORK THAT CAN **SCALE** WITH OUR COMPANY...

WITH LITTLE MAINTENANCE...

AND WITH OUR SMALL TEAM...

WHAT WE'LL COVER

- > WHAT IS KUBERNETES?
- > AUTOSCALING IN ACTION
- > HOW CAN I TRY IT OUT?

WHAT IS KUBERNETES?

KUBERNETES IS A PORTABLE, EXTENSIBLE OPEN-SOURCE PLATFORM FOR MANAGING CONTAINERIZED WORKLOADS AND SERVICES, THAT FACILITATES BOTH DECLARATIVE CONFIGURATION AND AUTOMATION.

KUBERNETES CLUSTERS

Kubernetes Architectural Overview

Pod

Pod

SCHEDULE

Pod

kubelet

kubelet

Node A

Node B
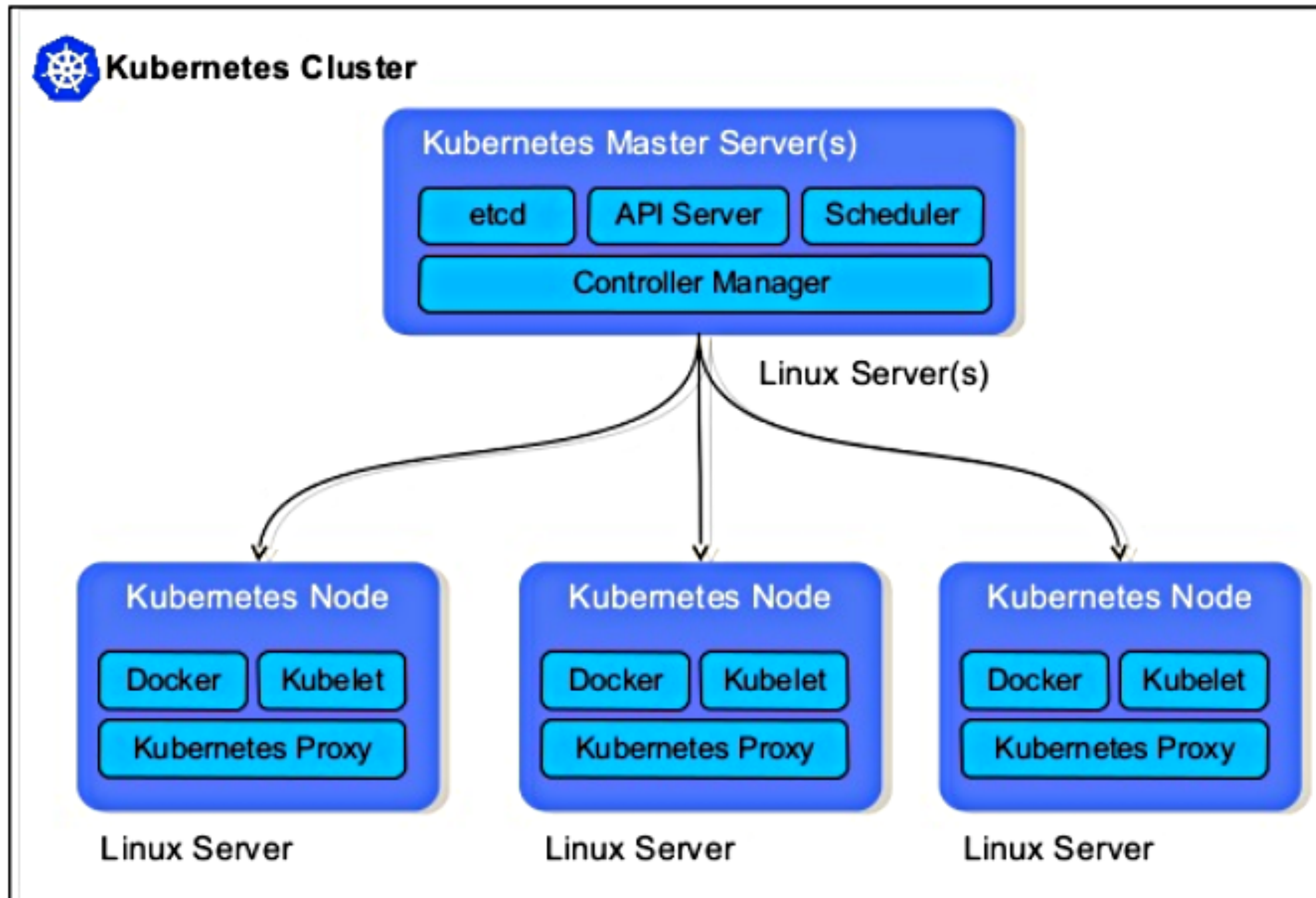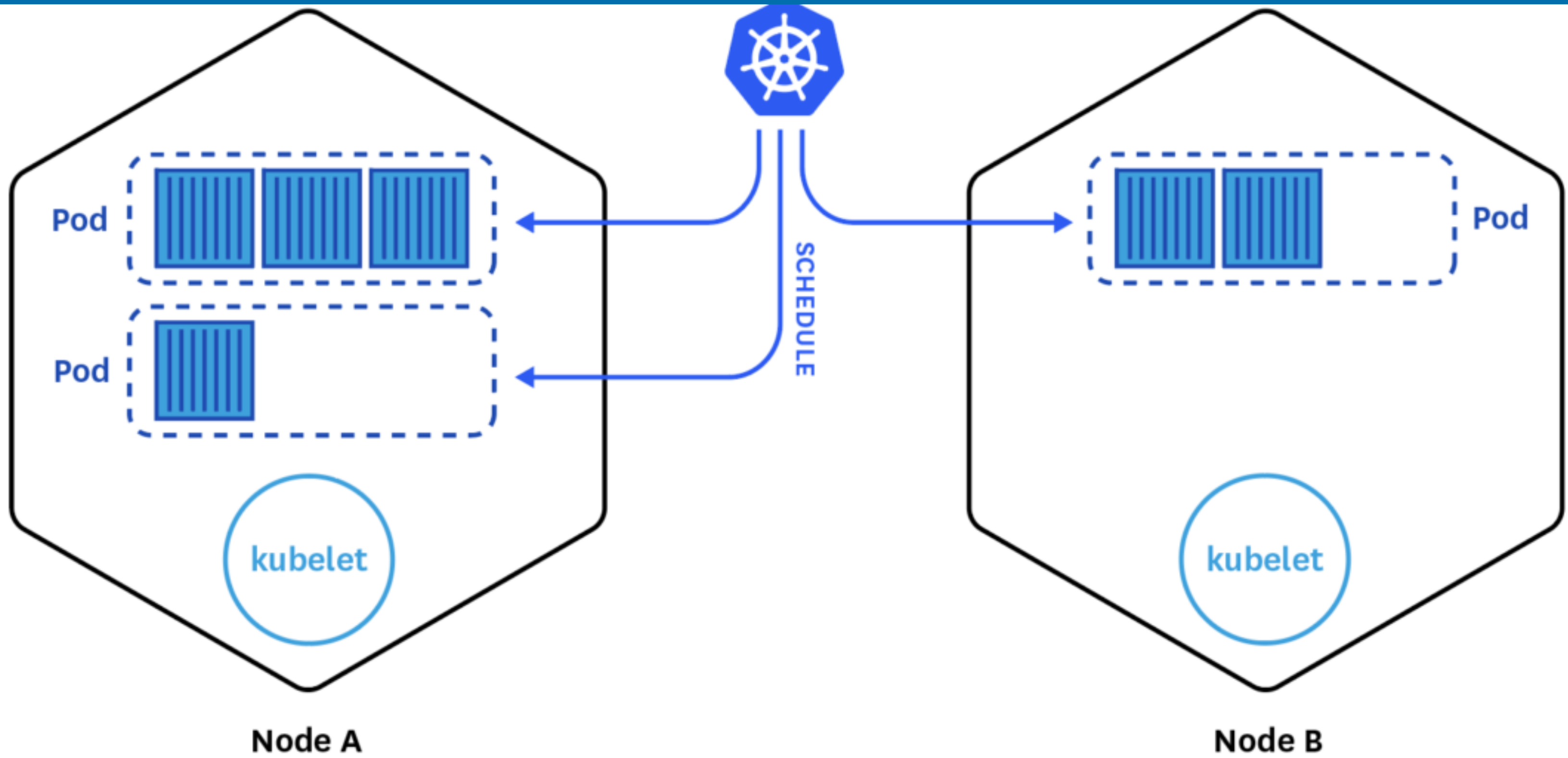
WHY IS KUBERNETES EXCITING?

BATTLE TESTED

DECLARATIVE: CONFIGURATION AS CODE!

SUPPORTED BY ALL MAJOR CLOUD PROVIDERS

HEALTHY OPEN SOURCE COMMUNITY

IS IT OVERKILL?

AUTOSCALING IN ACTION

WHAT DO WE HAVE TO DO?

> CONTAINERIZE OUR APP WITH DOCKER

> CONFIGURE A KUBERNETES CLUSTER

> WRITE A DEPLOYMENT

> WRITE A SERVICE

> WRITE A HORIZONTAL POD AUTOSCALER

> DEPLOY OUR APP TO THE CLUSTER!

```ruby
require 'sinatra'

get '/' do
  find_match_for_dog # some computationally expensive job
  "Your dog's match was found by #{ENV['POD_NAME']}"
end
```

CONTAINERIZING OUR APP

WHAT'S A CONTAINER ANYWAY?

A CONTAINER IMAGE IS A LIGHTWEIGHT, **STAND-ALONE,** EXECUTABLE PACKAGE OF A PIECE OF SOFTWARE THAT INCLUDES EVERYTHING NEEDED TO RUN IT.

DO YOU REMEMBER SETTING UP YOUR FIRST DEVELOPMENT MACHINE?

> INSTALL PACKAGES

> INSTALL LANGUAGES, FRAMEWORKS, AND TOOLS

> INSTALL DEPENDENCIES

> INSTALL SERVICES (MYSQL, POSTGRES, REDIS, ELASTICSEARCH, ETC...)

ONCE EVERYTHINGS INSTALLED, IT'S PROBABLY NOT THE SAME AS YOUR PRODUCTION ENVIRONMENT

CONTAINERS RUN ANYWHERE AND HAVE EVERYTHING THEY NEED TO RUN

docker

WRITING OUR DOCKERFILE

```dockerfile
# Specify a base image
# In this case, we want one that has Ruby preinstalled
FROM ruby:2.3.7

# Specify our working directory
WORKDIR /app

# Copy files from our current location on our local machine
# to the app directory in the Docker container
ADD . /app

# Install our dependencies
RUN cd /app && \
    bundle install

# Expose a port
EXPOSE 4567

# Start our server when the Docker container is started
CMD ["ruby", "web.rb"]
```

NOW THAT WE'VE CONTAINERIZED OUR APP. WE'RE READY TO DEPLOY IT WITH KUBERNETES

WE'RE GOING TO WRITE SOME MANIFESTS THAT DESCRIBE WHAT OUR APPLICATION SHOULD LOOK LIKE

DEPLOYMENT

A DEPLOYMENT DECLARES YOUR DESIRED STATE OF AN APPLICATION, USUALLY AS A SERIES OF PODS AND HOW THEY SHOULD BE MANAGED

POD

A POD (AS IN A POD OF WHALES OR PEA POD) IS A GROUP OF ONE OR MORE CONTAINERS (SUCH AS DOCKER CONTAINERS), WITH SHARED STORAGE/NETWORK, AND A SPECIFICATION FOR HOW TO RUN THE CONTAINERS.

IN LAYMANS TERMS, PODS ARE UNITS OF WORKS OR REPLICAS OF AN APPLICATION. AND DEPLOYMENT DESCRIBES HOW TO CONFIGURE THEM.

```yaml
apiVersion: extensions/v1beta1
kind: Deployment
metadata:
  name: dog-tinder
spec:
  template:
    metadata:
      name: dog-tinder
      labels:
        service: dog-tinder
    spec:
      containers:
      - name: dog-tinder-container
        image: dog-tinder:latest
        resources:
          requests:
            cpu: 200m
        env:
        - name: POD_NAME
          valueFrom:
            fieldRef:
              fieldPath: metadata.name
      restartPolicy: Always
```

WHAT DOES IT DO?

```yaml
# How much resources should we allocate to this pod?
resources:
  requests:
    cpu: 200m
```

1/5TH OF A CPU CORE

SERVICE

A SERVICE ALLOWS OTHER THINGS TO ACCESS YOUR DEPLOYMENT. IN OUR CASE WE WANT TO VIEW OUR APP FROM A BROWSER!

IN OTHER CASE YOU MIGHT NEED TO CONFIGURE A SERVICE SO THAT YOUR APP COULD CONNECT TO A DATABASE

```yaml
apiVersion: v1
kind: Service
metadata:
  labels:
    service: dog-tinder
  name: dog-tinder-service
spec:
  type: NodePort
  ports:
  - name: dog-tinder-port
    port: 4000
    protocol: TCP
    targetPort: 4567
    nodePort: 30000
  selector:
    service: dog-tinder
```

HORIZONTAL POD AUTOSCALER

```yaml
apiVersion: autoscaling/v1
kind: HorizontalPodAutoscaler
metadata:
  name: dog-tinder-hpa
spec:
  maxReplicas: 10
  minReplicas: 1
  scaleTargetRef:
    apiVersion: extensions/v1beta1
    kind: Deployment
    name: dog-tinder
  targetCPUUtilizationPercentage: 50
```

targetCPUUtilizationPercentage: 50

50% * 200M = 100M

IF WE USE OUR FULL 200M. THE HPA WILL TRY TO SCALE TO 2 PODS

DEMO

HOW CAN I LEARN MORE?

AND IF YOU WANT TO SEE ALL THE STEPS FOR THE THING I JUST DID...

HTTP://WWW.NOOBERNETES.IO

CHECK IT OUT, FILE AN ISSUE IF ITS CONFUSING AND I'LL UPDATE IT!

THE BEST WAY TO LEARN IS TO DO

QUESTIONS?