



SHADOW READER

SERVERLESS LOAD TESTING FOR REPLAYING TRAFFIC

Yuki Sawa @yukisww

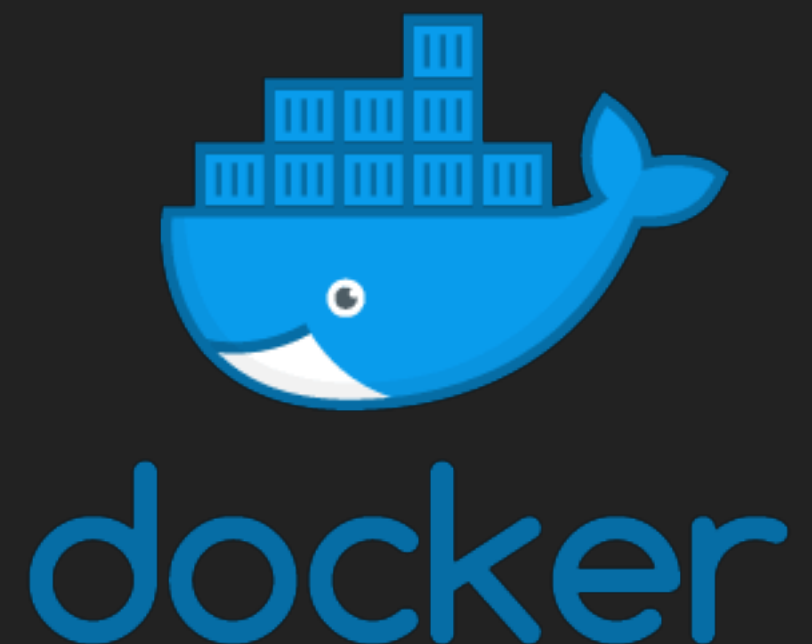
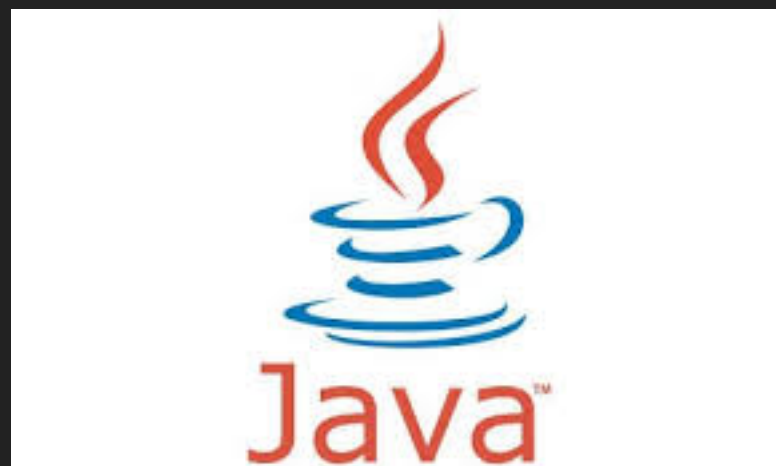
Software Engineer

edmunds.com

github.com/edmunds/shadowreader



edmunds

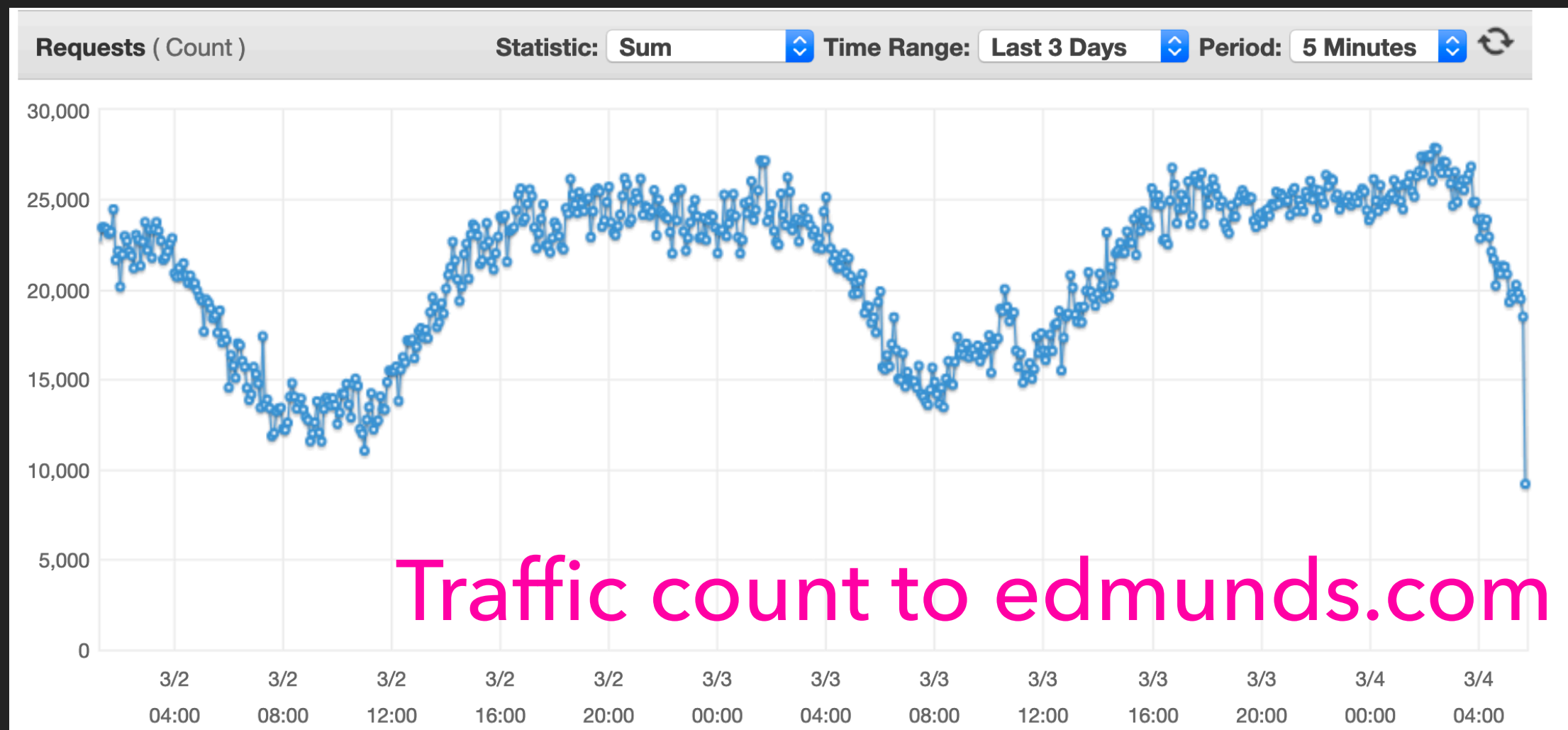


SUMMARY

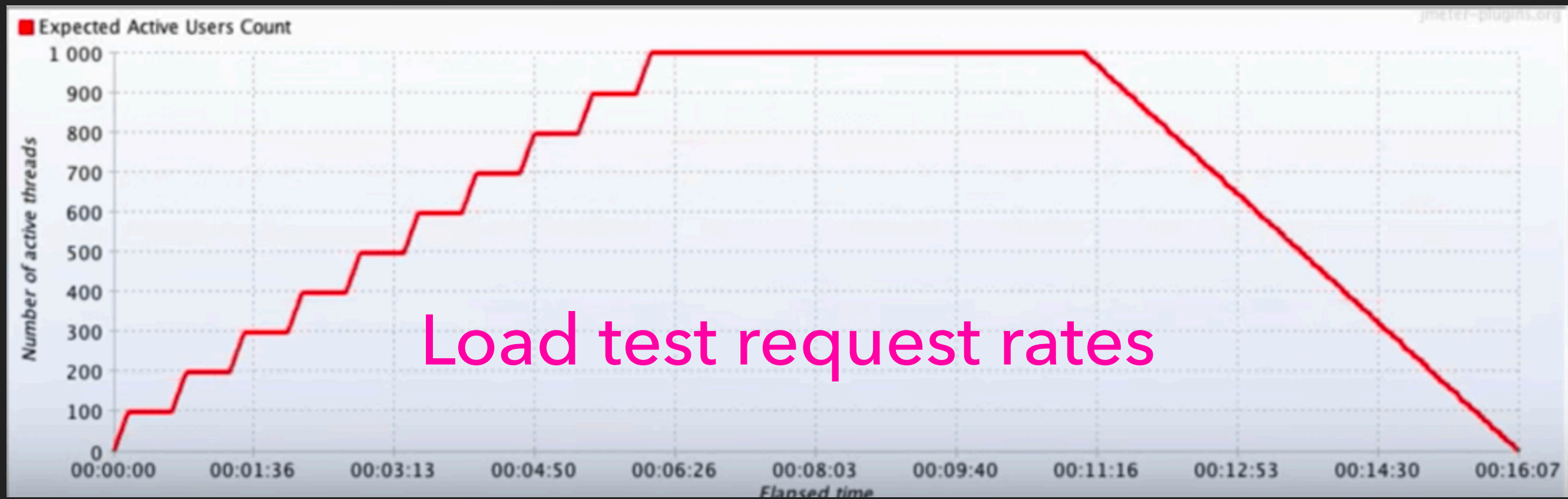
- ▶ Challenges of load testing
- ▶ How we tried to solve it
- ▶ How it solved an incident
- ▶ Architecture

HARD PARTS OF LOAD TESTING

► Need real request rates



► Synthetic load test



▶ Need realistic test URLs

- ▶ [edmunds.com/used-cars](https://www.edmunds.com/used-cars)

- ▶ [edmunds.com/used-hondas](https://www.edmunds.com/used-hondas)

- ▶ [edmunds.com/ford](https://www.edmunds.com/ford)

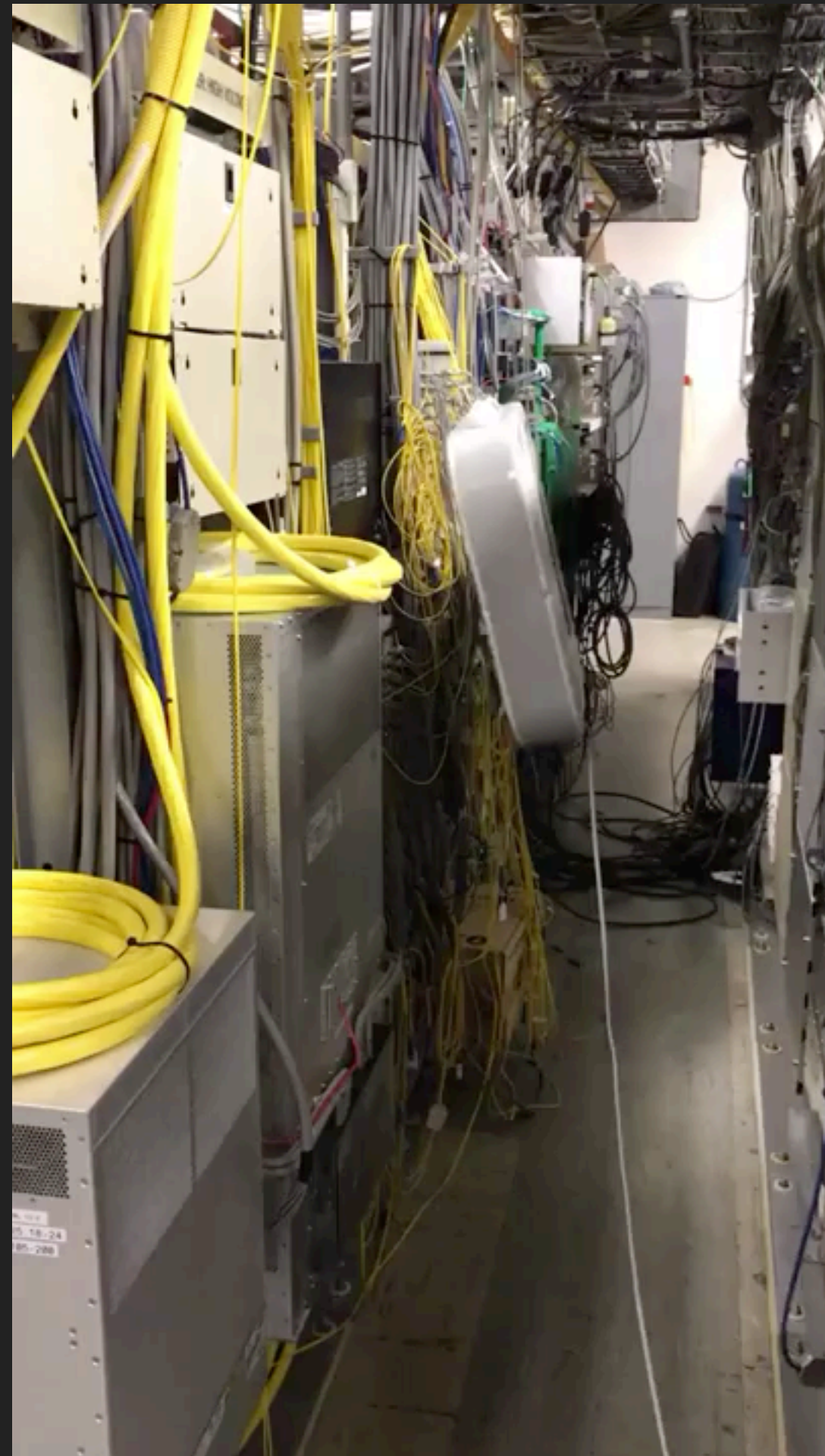
- ▶ [edmunds.com/suv](https://www.edmunds.com/suv)

/inven/srp?e2e=true&__mode=noss&test-
zip=53545&radius=100&inventoryType=used&make=mercury&mod
el=mariner&trim=premier&year=2008-2008&extcolor=%22Black
+Clearcoat%7C0%2C0%2C0%22&price=6500-6500&vin=4M2CU97128
KJ10613&fassignment=venom-used-lead-form-
srp%3Achal2%7Cnpf-transparent-pricing%3Actrl&enable-
feature=amp,gtm,inlineCritical,sentient,spaLinks,spaRule
s,sentry,wtf,wtfCache,CORE-67-PreProdCore,TRAF-2836-
ContextualLinks,TRAF-3125-
NoCarMakesUnderResearch,UPF-1306-AMP-SEO,UPF-1333-
NationwideRegionalDelivery,TRAF-3233-
RemoveCarsForSale,ADS-2377-
DCOSRPNative,homeAutoComplete,vdpsavesubnav,savecoresubn
av&disable-
feature=loadFeatureFlagsFromS3,ads,ampAnalytics,showErro
rs,generateErrorProdTemplate,testSpa,modelLru,wtfShowErr
ors,trackerWrapper,nativeFedTest,ADS-1541-
DCOPricingNative,CBP-1252-SRPCardLabel,ADS-1657-
Fluid,pfS3Photo,npf-show-checkboxes,CORE-602-remove-
image-carousel,CORE-533-RankingsGenerations


```
/gateway/graphql/?
query=query%20(%24makeSlug%3A%20String!
%2C%20%24modelSlug%3A%20String!%2C%20%24year%3A%20Int!)
%20%7B%0A%20%20allVehicles(makeSlug%3A%20%24makeSlug)
%20%7B%0A%20%20%20models(modelSlug%3A%20%24modelSlug)
%20%7B%0A%20%20%20%20modelYears(year%3A%20%24year)
%20%7B%0A%20%20%20%20%20segmentRatings%20%7B%0A
%20%20%20%20%20%20rank%0A%20%20%20%20%20%20%20
%20%20%20%20slugRankedSubmodel%0A%20%20%20%20%20%20%20
%20%20editorialSegment%20%7B%0A%20%20%20%20%20%20%20%20
%20%20%20%20edmundsTypeCategory%0A%20%20%20%20%20%20%20%20
%20%20%20%20id%0A%20%20%20%20%20%20%20%20%20%20display
Name%0A%20%20%20%20%20%20%20%20%20%20segmentR
atings%20%7B%0A%20%20%20%20%20%20%20%20%20%20%20%20
%20rating%0A%20%20%20%20%20%20%20%20%20%20%20%7D%0A%20%
%20%20%20%20%20%20%20%20%20%7D%0A%20%20%20%20%20%20%20
%7D%0A%20%20%20%20%20%20%20%7D%0A%20%20%20%20%7D%0A%20%7
D%0A%7D%0A&variables=%7B%22makeSlug%22%3A%22bmw%22%2C%22
modelSlug%22%3A%225-series%22%2C%22year%22%3A2019%7D
```

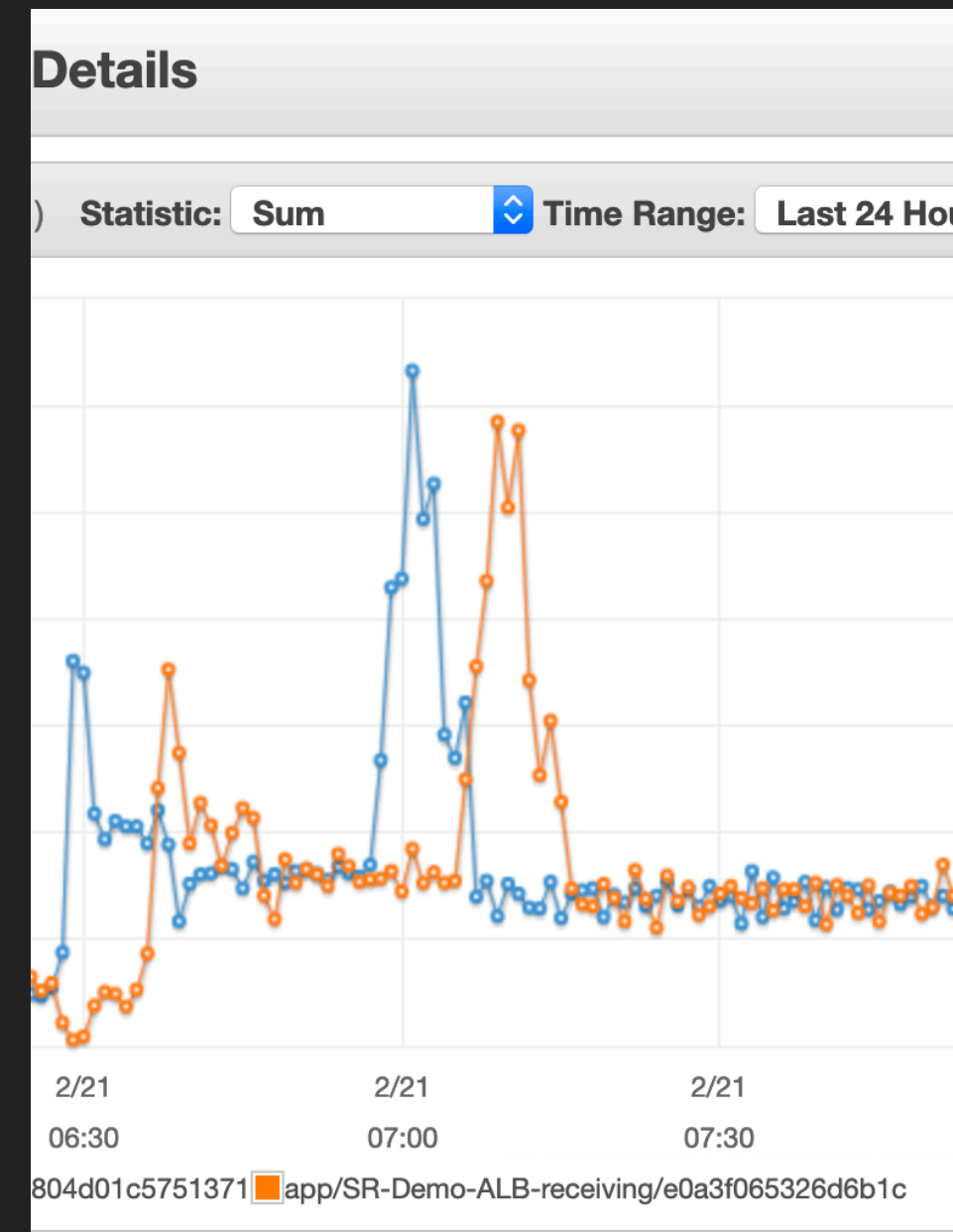

CHALLENGES OF LOAD TESTING

- ▶ Distributed load tests
 - ▶ Maintenance 
 - ▶ Load test configs
 - ▶ Boot up scripts
 - ▶ CPU/MEM allocation
 - ▶ Server costs 
- ▶ Load test takes time to start up

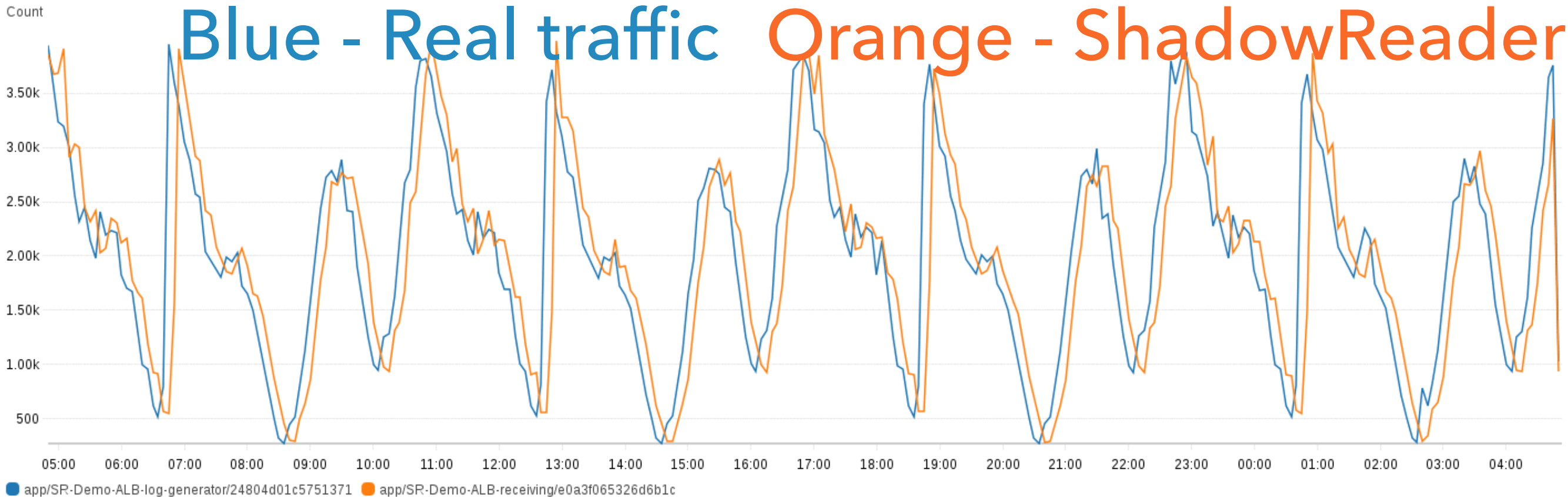


SHADOWREADER

- ▶ From a Hackathon in November
 - ▶ Used in prod in January
- ▶ Replay URLs
- ▶ Replay request rate
- ▶ Serverless
- ▶ AWS Lambda



Request counts



SHADOWREADER – WHAT IS IT?

- ▶ Simulate production traffic in QA
 - ▶ Pre-prod canary deploys
- ▶ Replay peak traffic hour

THE CHRISTMAS EVE MEMORY LEAK

- ▶ Memory leak caused high errors rates in prod
- ▶ Couldn't be reproduced in QA
- ▶ ShadowReader used to solve it

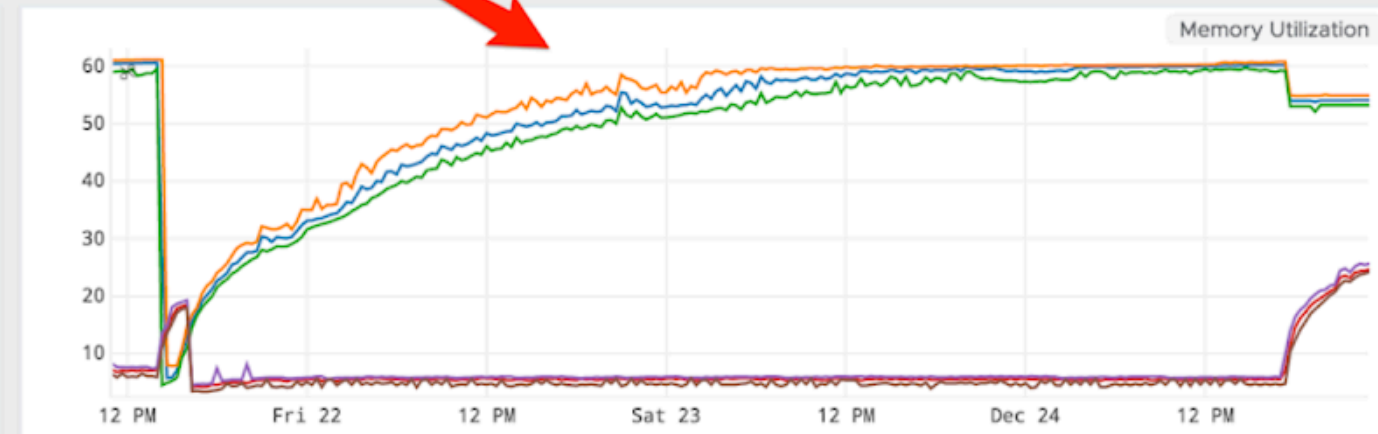
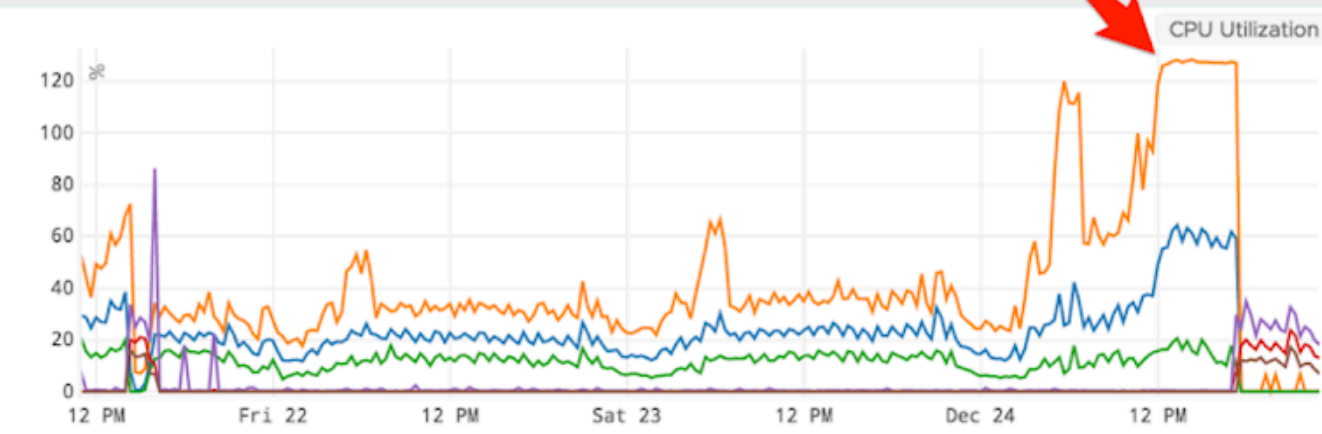
CHRISTMAS EVE INCIDENT

- ▶ December 24th, 2017
 - ▶ Memory leak causes high error rates and latency
 - ▶ On-call alerted 🤔
 - ▶ Resolved by Docker Orchestration engine (ECS)

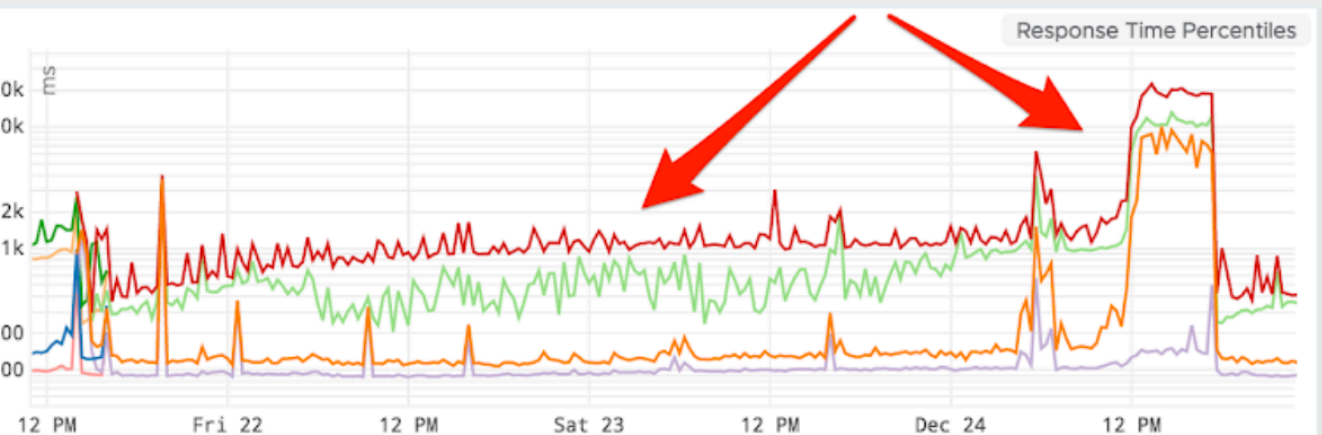
1. Memory grows over a course of a few days

2. CPU spikes at the end

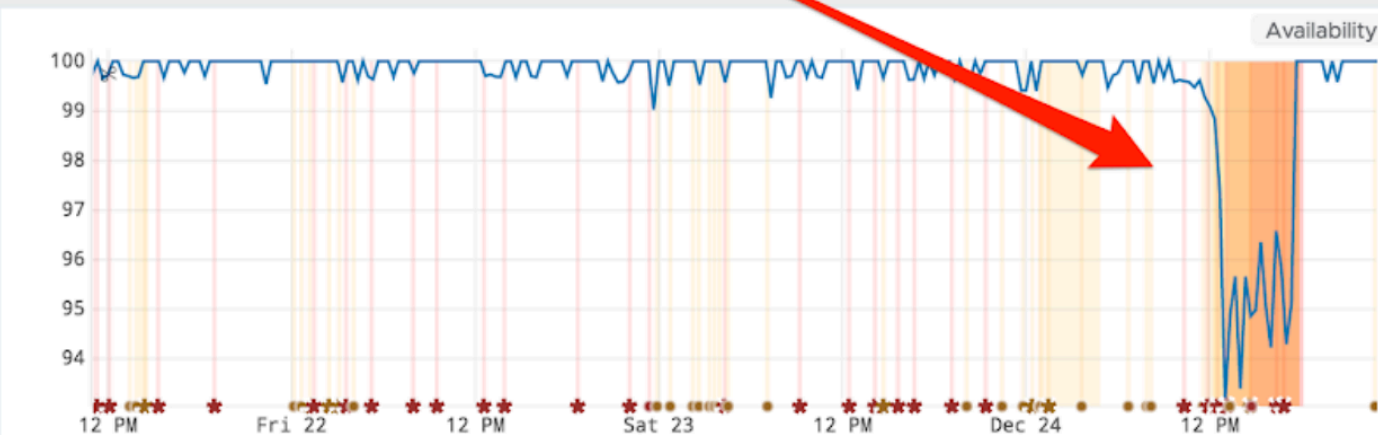
Hardware ▾



Customer response times are increasing the whole time.



Finally at the end availability starts to drop

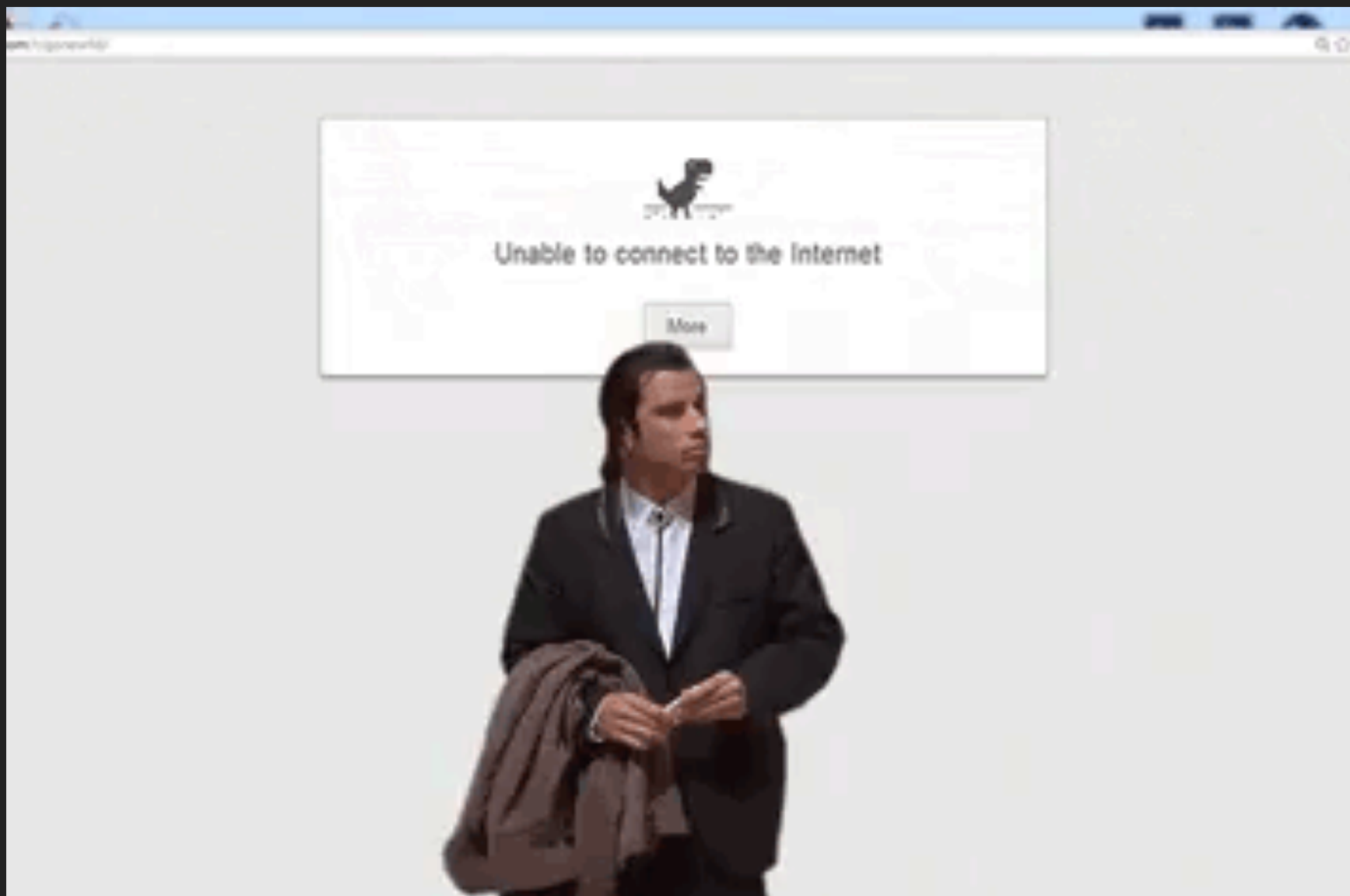


EDMUNDS INFRASTRUCTURE

- ▶ Docker on ECS
- ▶ Canary releases in prod and qa
- ▶ AutoScaling
 - ▶ => Masked memory leak!

INVESTIGATION

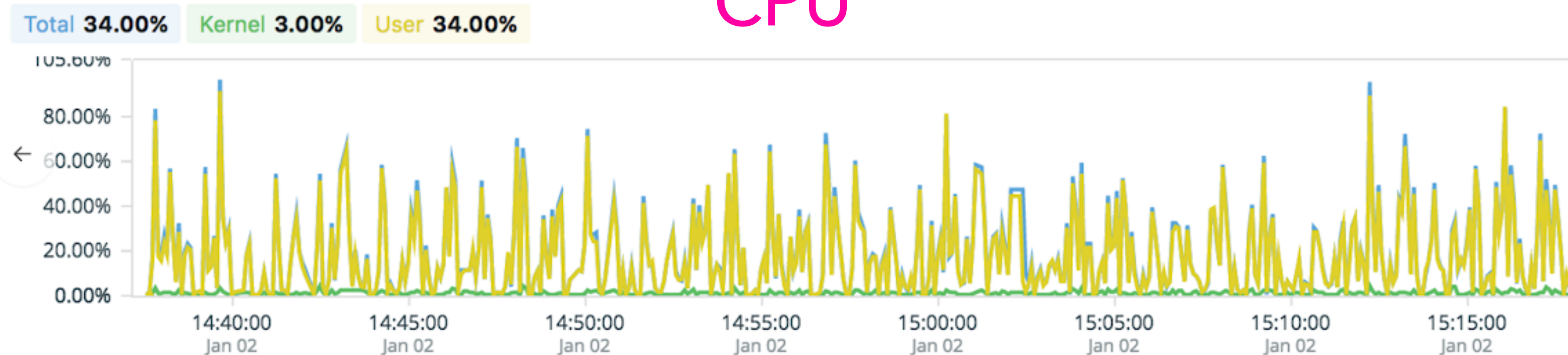
► QA doesn't have a memory leak!



PROD

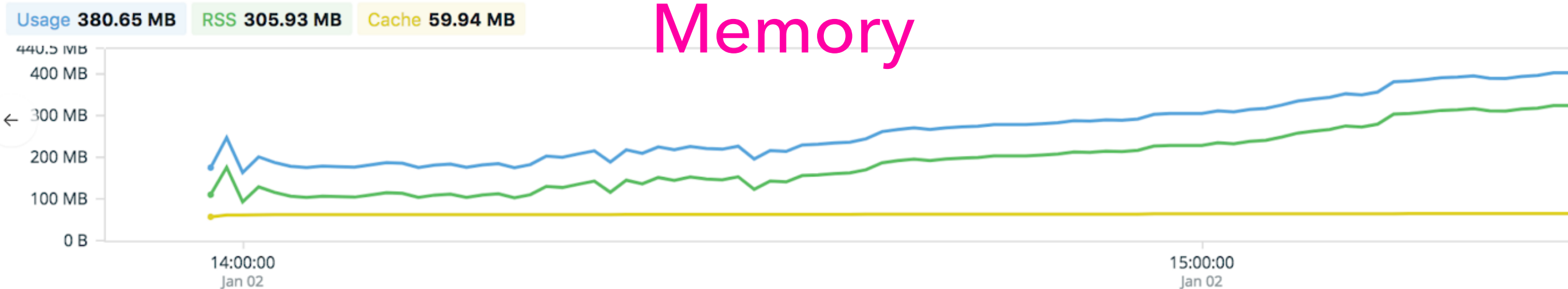
CPU

CPU



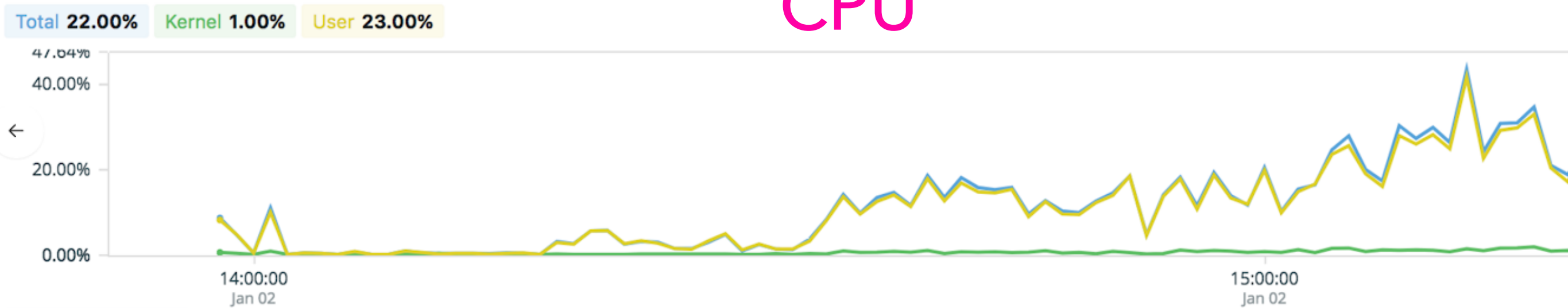
Memory (Limit: 2 GB)

Memory



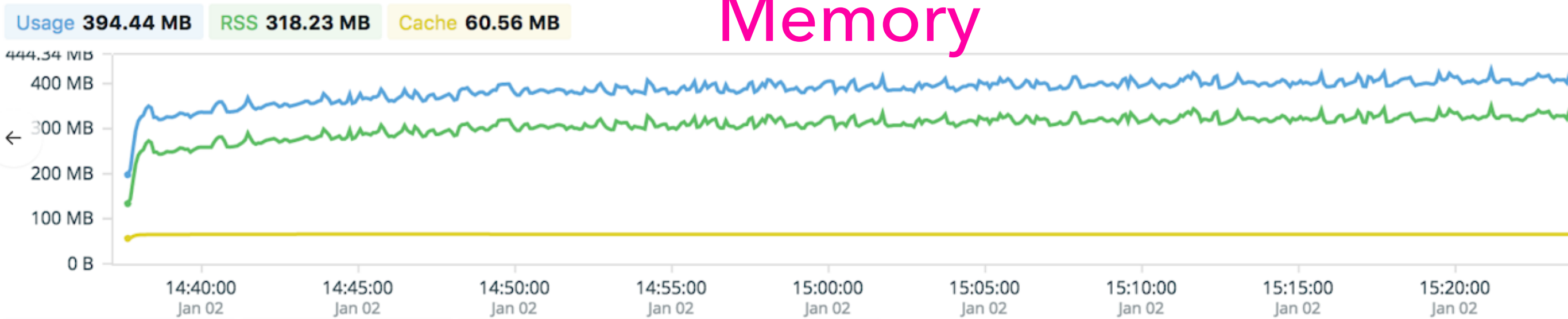
QA

CPU



CPU

Memory (Limit: 2 GB)



Memory

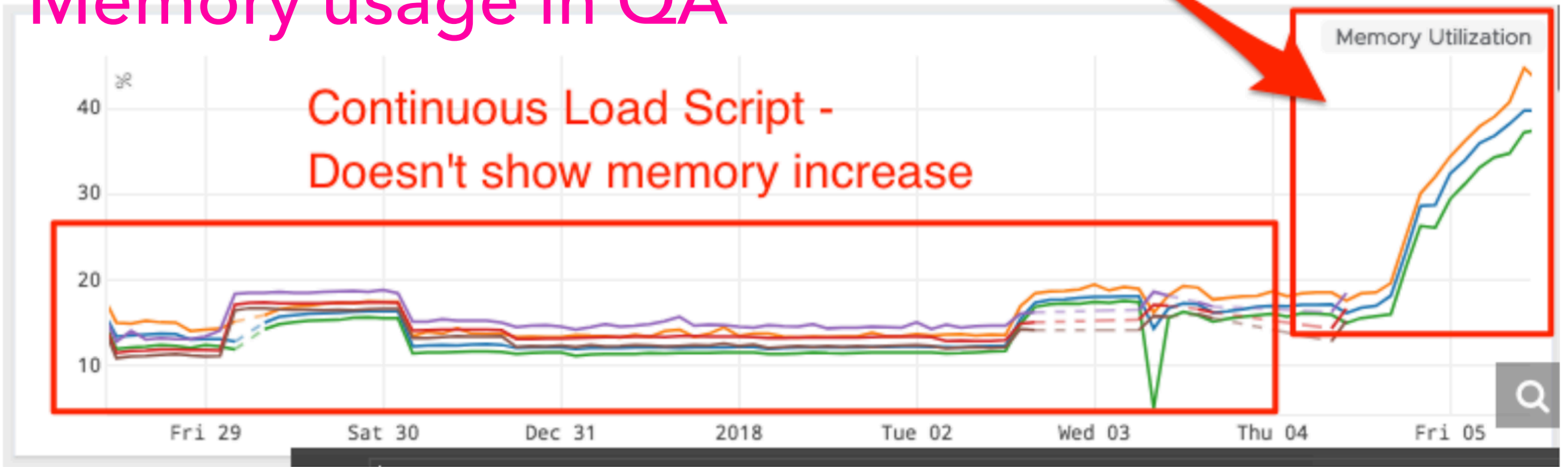
INVESTIGATION

- ▶ Maybe load test can't recreate it in QA?
- ▶ Synthetic load tests in QA
 - ▶ Using URLs generated by scripts or by hand
 - ▶ Static request rates



Shadow Reader test in QA-11 -
Reflects memory leak pattern
similarly observed in prod

Memory usage in QA



- ▶ Introduce ShadowReader
- ▶ Saw results immediately

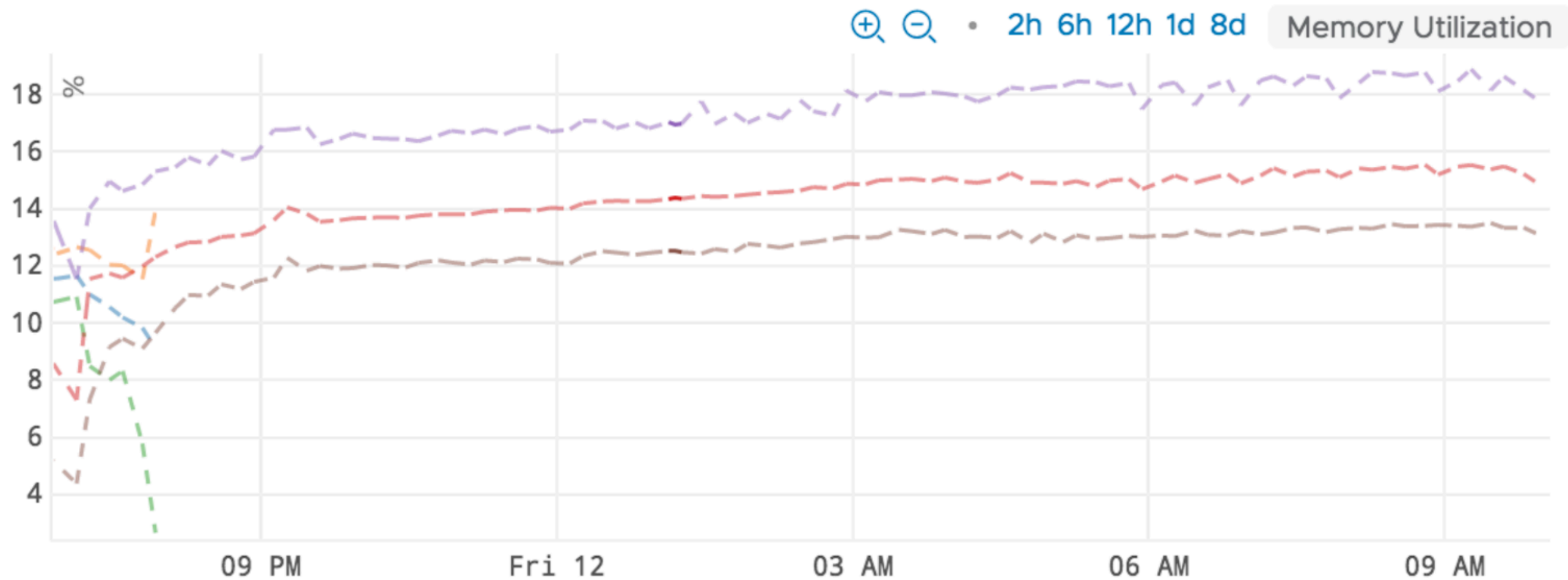
THE CAUSE

- ▶ Point ShadowReader to local
- ▶ 400MB of metadata
- ▶ Server caching for all users
 - ▶ Cache was never being used!

THE SOLUTION

- ▶ Synthetic load test didn't test with enough URLs/throughput to simulate enough users
- ▶ Replay traffic
 - ▶ => generated enough unique meta data in the cache

- ▶ Disabled server side caching
- ▶ Memory nice and even 🙌 😊



SHADOWREADER ARCHITECTURE

- ▶ Tools and AWS Services
- ▶ ShadowReader features
 - ▶ Replay traffic
 - ▶ Serverless
- ▶ Design and architecture

TOOLS

- ▶ Serverless framework
- ▶ Python 3



AWS SERVICES

- ▶ AWS Lambda
- ▶ S3
- ▶ Elastic Load Balancers
- ▶ CloudWatch Events

REPLAY LOAD TEST

- ▶ Parses production access logs and replays it
- ▶ Replay request rate and URLs

REPLAY LOAD TEST

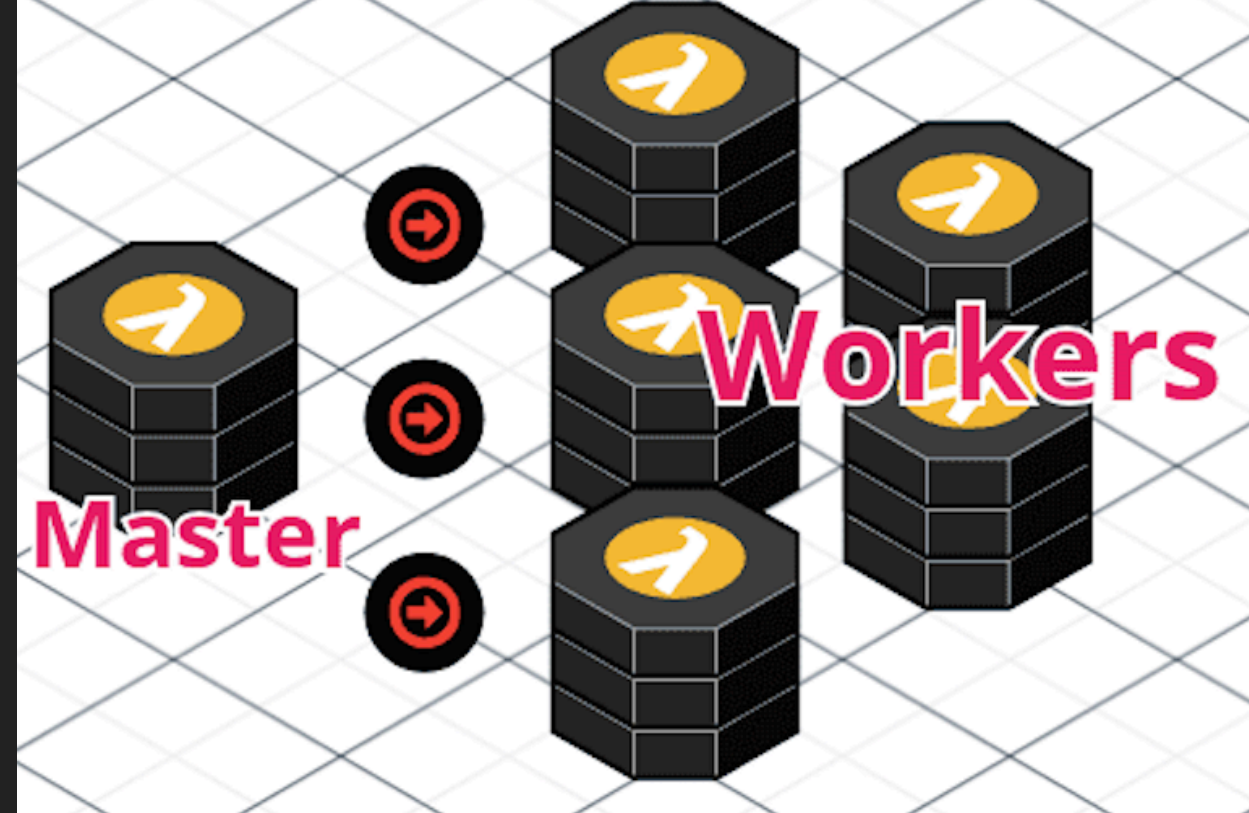
- ▶ Live replay or Past replay
- ▶ Replay headers
 - ▶ User-Agent or True-Client-IP

SERVERLESS

- ▶ Easy to scale
- ▶ Provision on demand
- ▶ Scale to 50k reqs / minute
- ▶ Cheap 
- ▶ Pay only for what you use
- ▶ \$1000+ / month → \$100 / month

SERVERLESS

- ▶ Achieved by
 - ▶ High concurrency
 - ▶ 100 requests / Worker Lambda
 - ▶ 256MB MEM / Lambda
 - ▶ No maintenance, fast start up




```
[  
  {  
    "uri": "/post1",  
    "req_method": "GET",  
    "timestamp": "2019-02-21 03:39:00+00:00",  
    "user_agent": "Mozilla/5.0 Firefox/7.0.1",  
    "IP": "1.2.3.4"  
  }  
]
```

- ▶ Test data partitioned into minute intervals
- ▶ 1 minute of traffic == list of URLs from that minute
- ▶ 1 hour of traffic == 60 jsons
- ▶ An array of URLs for each minute of the day
- ▶ All URL data stored on S3

OTHER FEATURES

- ▶ Plugin system - choose live or past replay
- ▶ Support for replaying
 - ▶ Application/Elastic Load Balancer
- ▶ Ramp traffic by % value

ARCHITECTURE

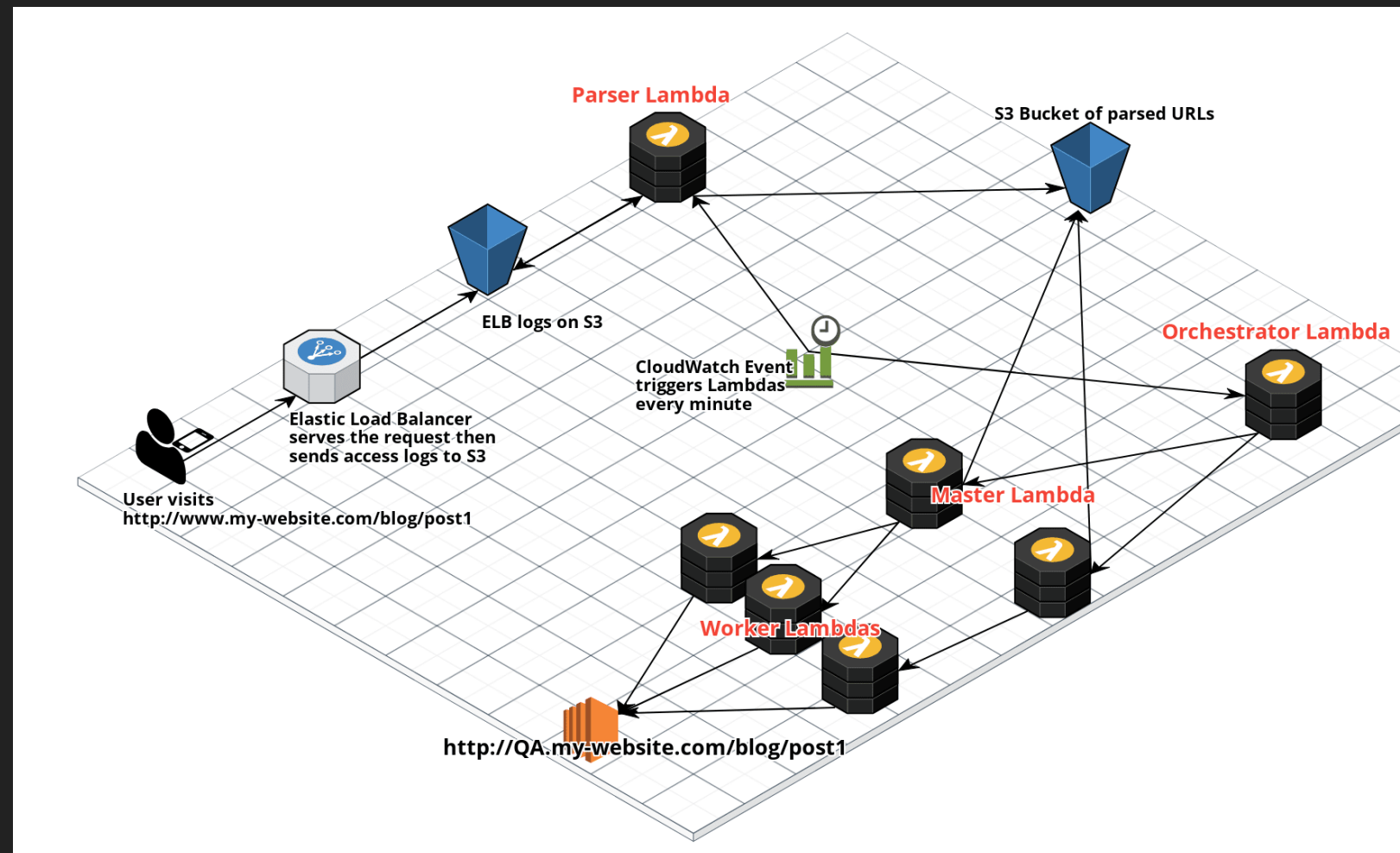
► 4 Lambdas

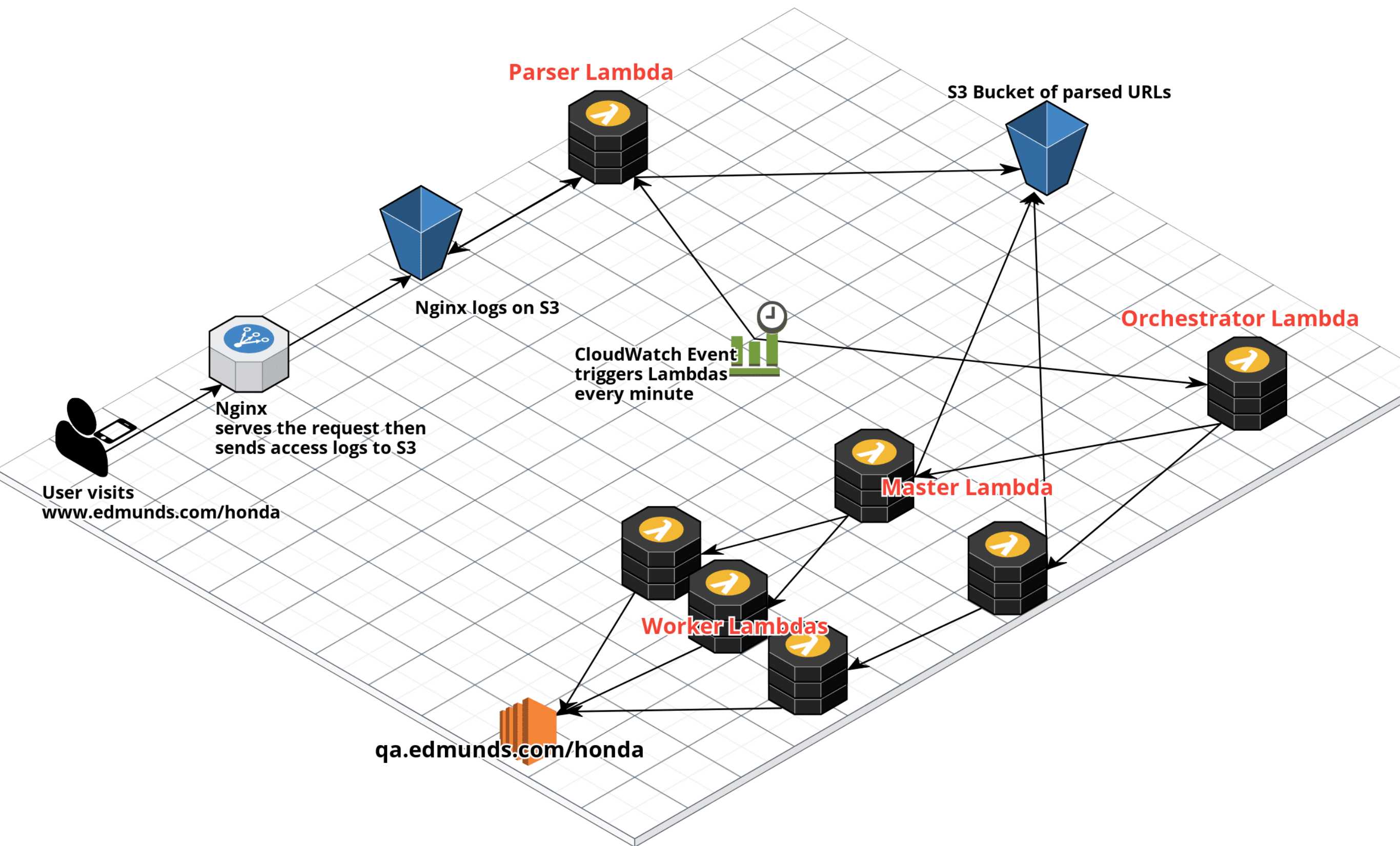
► Parser

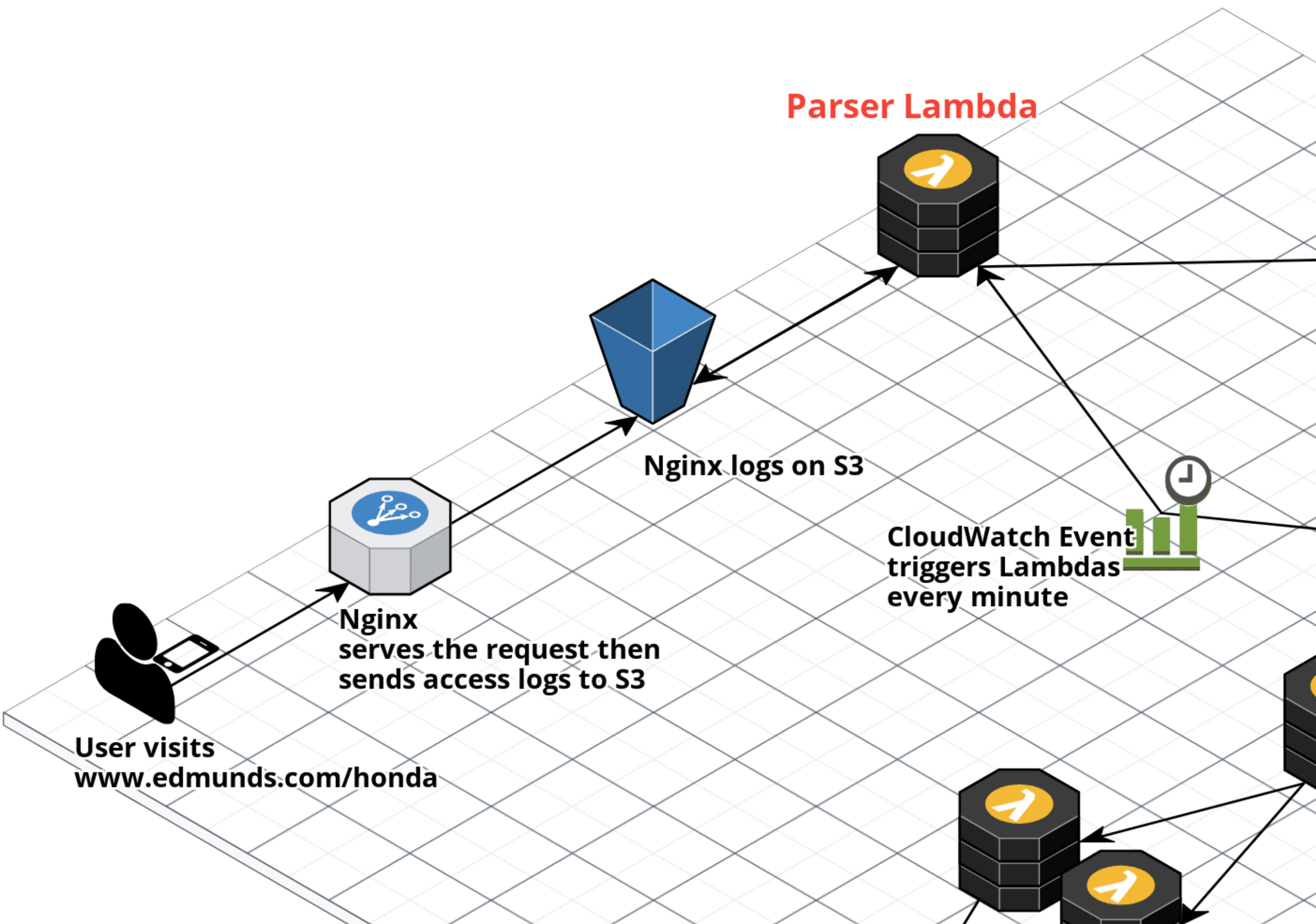
► Orchestrator

► Master

► Worker







Parser Lambda

S3 Bucket of parsed URLs

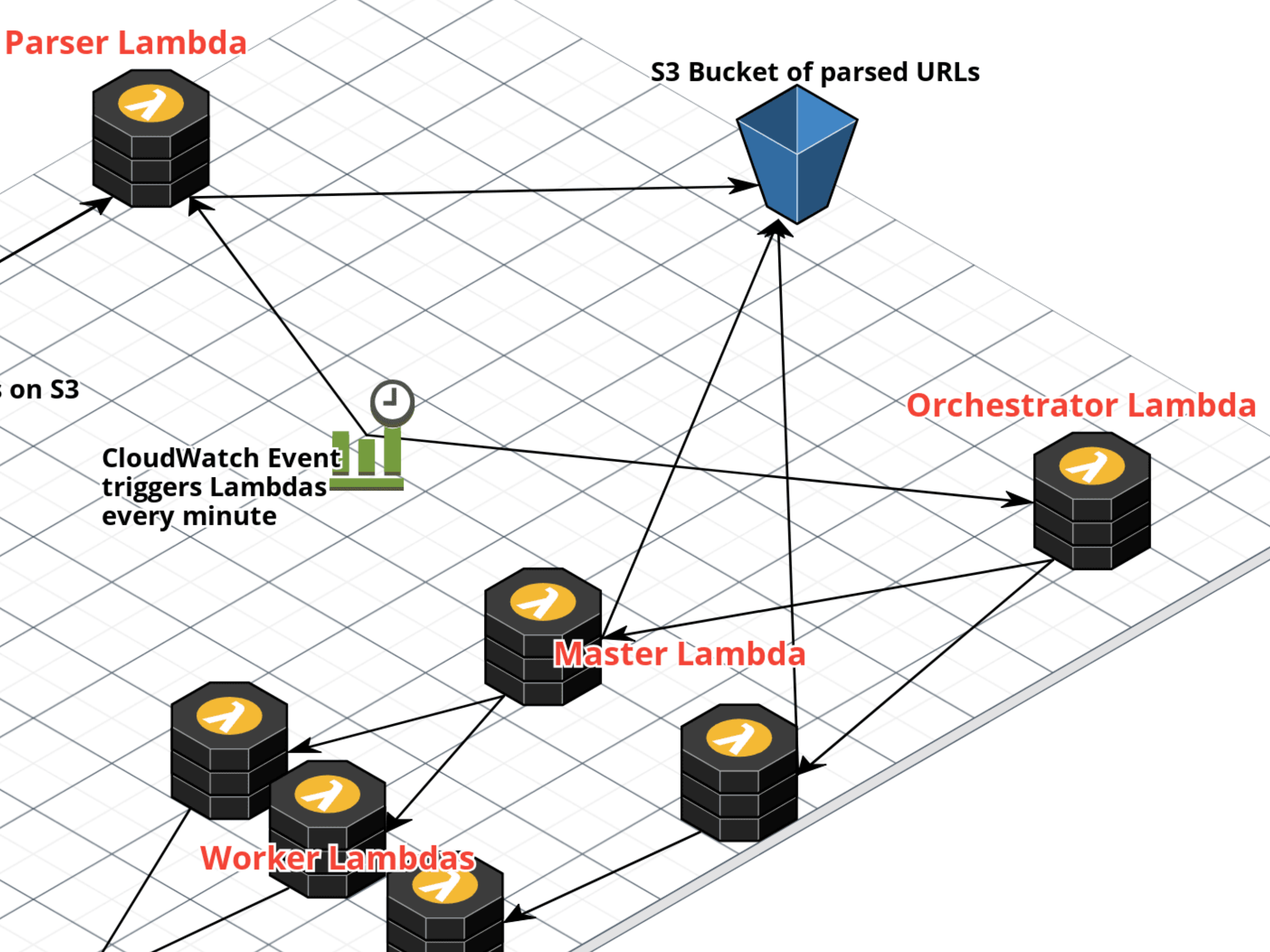
on S3

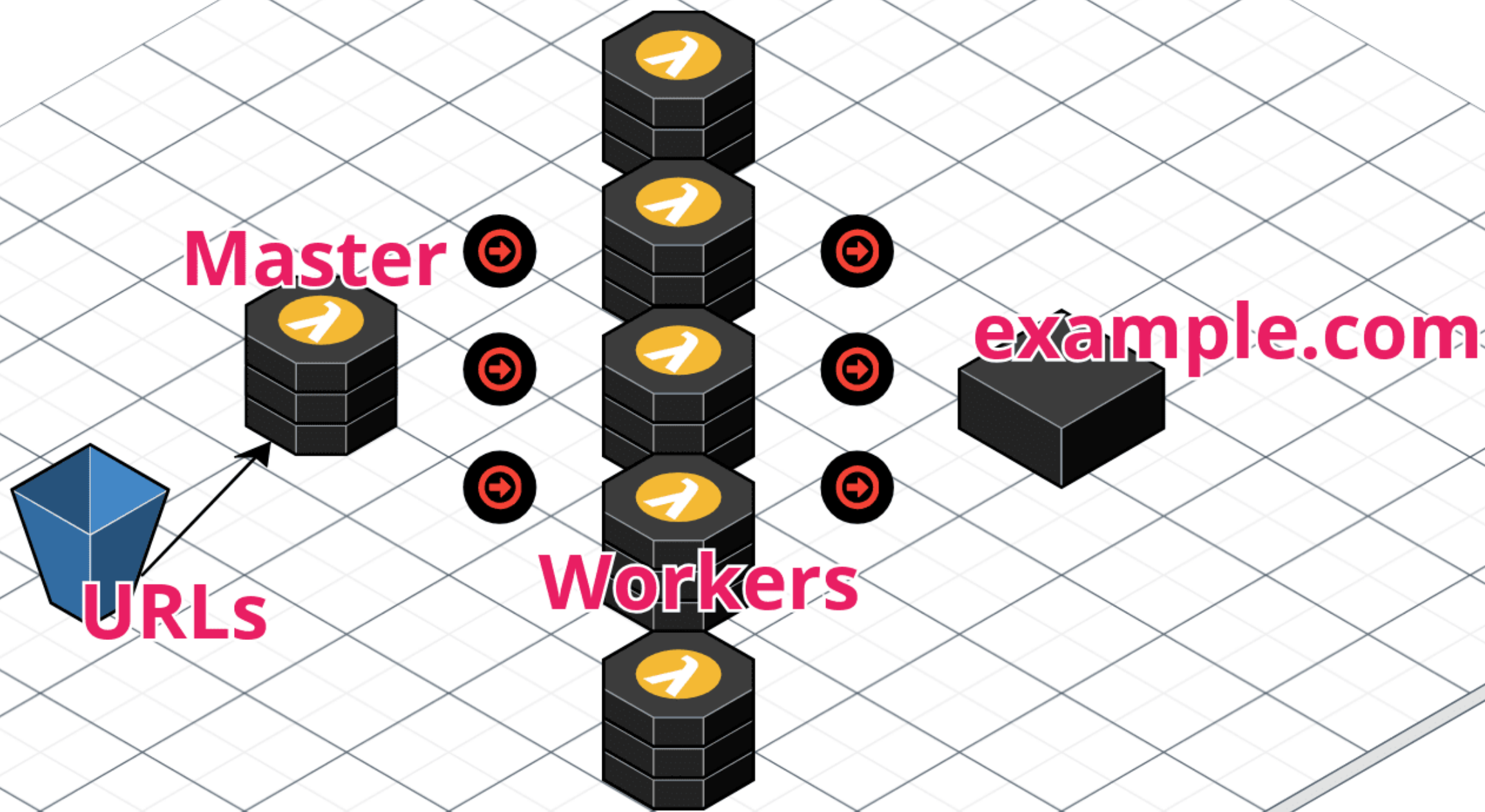
Orchestrator Lambda

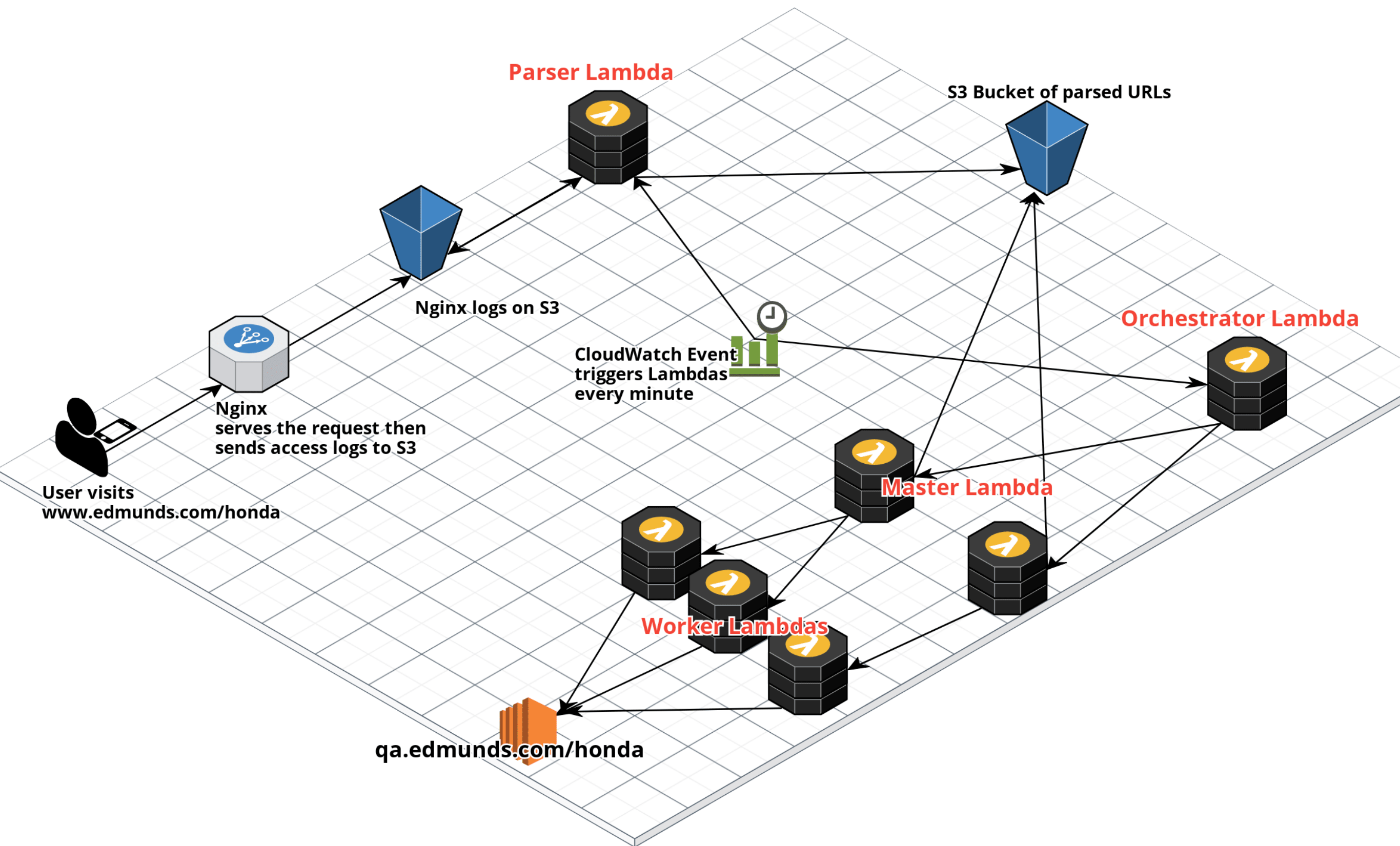
**CloudWatch Event
triggers Lambdas
every minute**

Master Lambda

Worker Lambdas







DEMO

- ▶ github.com/edmunds/shadowreader
- ▶ Welcoming contributions 😊
 - ▶ Replay HAProxy, other LBs, etc.
- ▶ Feedback / suggestions welcomed!!

DEMO

- ▶ github.com/edmunds/shadowreader
- ▶ Serverless framework
 - ▶ `npm install serverless`
 - ▶ `sls deploy`

```
Serverless: Removing old service artifacts from S3...
(p36) [rc:..ast-replay-poc/shadowreader]$ sls deploy --stage past-replay --region us-west-1 --aws-profile poc
Serverless: Generated requirements from /Users/rc/repos/sr-past-replay-poc/shadowreader/requirements.txt in /Users/rc/repos/sr-past-replay-poc/shadowreader/requirements.txt...
Serverless: Installing requirements from /Users/rc/repos/sr-past-replay-poc/shadowreader/.serverless/requirements/requirements.txt...
Serverless: Running ...
Serverless: Packaging service...
Serverless: Excluding development dependencies...
Serverless: Injecting required Python packages to package...
Serverless: Uploading CloudFormation file to S3...
Serverless: Uploading artifacts...
Serverless: Uploading service sr.zip file to S3 (1.74 MB)...
Serverless: Validating template...
Serverless: Updating Stack...
Serverless: Checking Stack update progress...
.....
Serverless: Stack update finished...
Service Information
service: sr
stage: past-replay
region: us-west-1
stack: sr-past-replay
resources: 14
api keys:
  None
endpoints:
  None
functions:
  orchestrator-past: sr-past-replay-orchestrator-past
  producer: sr-past-replay-producer
  consumer-worker: sr-past-replay-consumer-worker
  consumer-master-past: sr-past-replay-consumer-master-past
layers:
  None
Serverless: Removing old service artifacts from S3...
```

Update available 1.37.1 → 1.38.0
Run `npm i serverless` to update

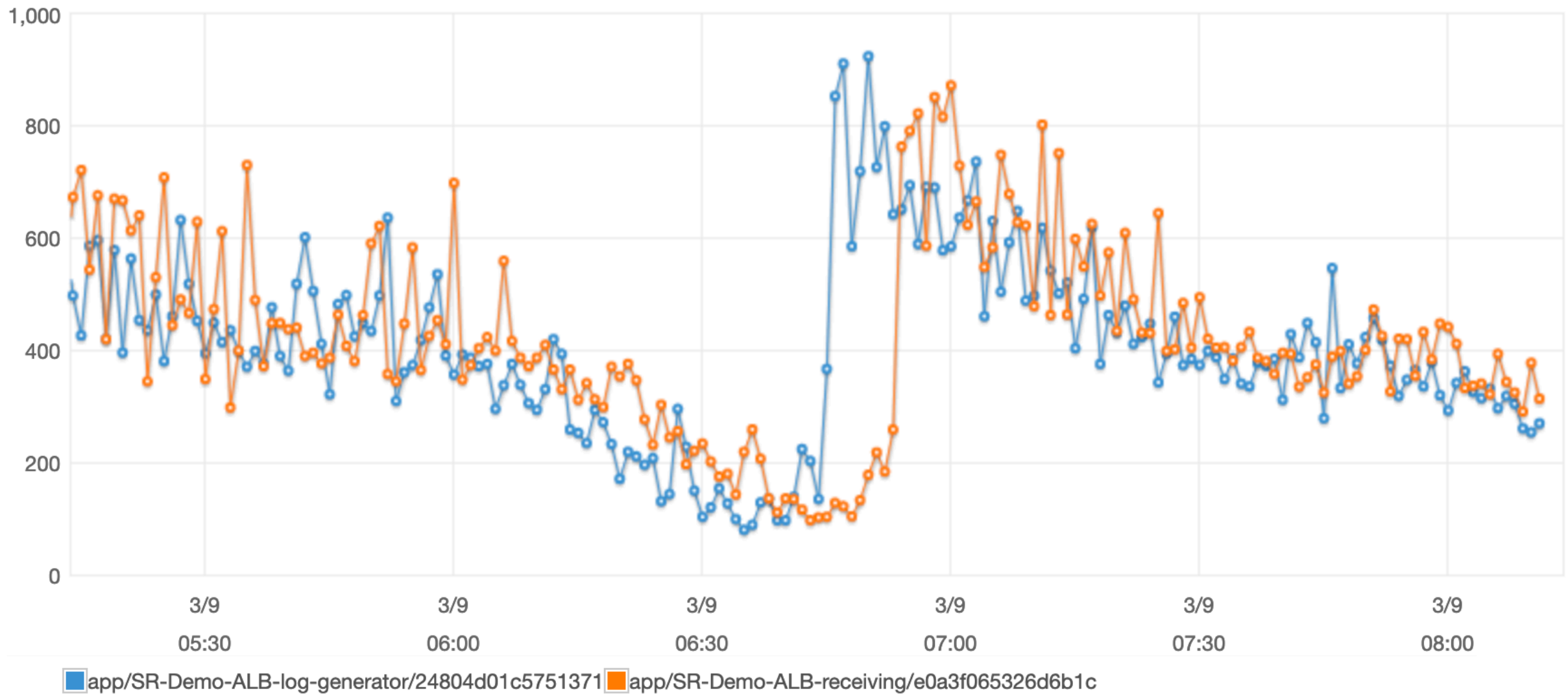
CloudWatch Monitoring Details

HTTP fixed response count (Count)

Statistic: **Sum**

Time Range: **Last 3 Hours**

Period: **1 Minute**



Close

THANK YOU

AND SPECIAL THANKS TO EVERYONE THAT HELPED

CARLOS MACASAET

EMIL NDREU

SHARATH GOWDA

INAYATULLAH BHOLAT

HABIB KHAN

LELAND SO

DENISE NGAI

ILANA GERSHTEYN

MONICA AIMA

NATALIA HRYSHCHANKOVA

PETER PURWANTO

[github.com/edmunds/
shadowreader](https://github.com/edmunds/shadowreader)

@yukisww