Scaling APIs from 0 to 60k RPM

IN A FAST GROWING STARTUP

PyParis - 2018/11/14

Who Am I?

Jean-Baptiste Aviat

CTO & Co-founder of sqreen.io
Former hacker at Apple (Red Team)

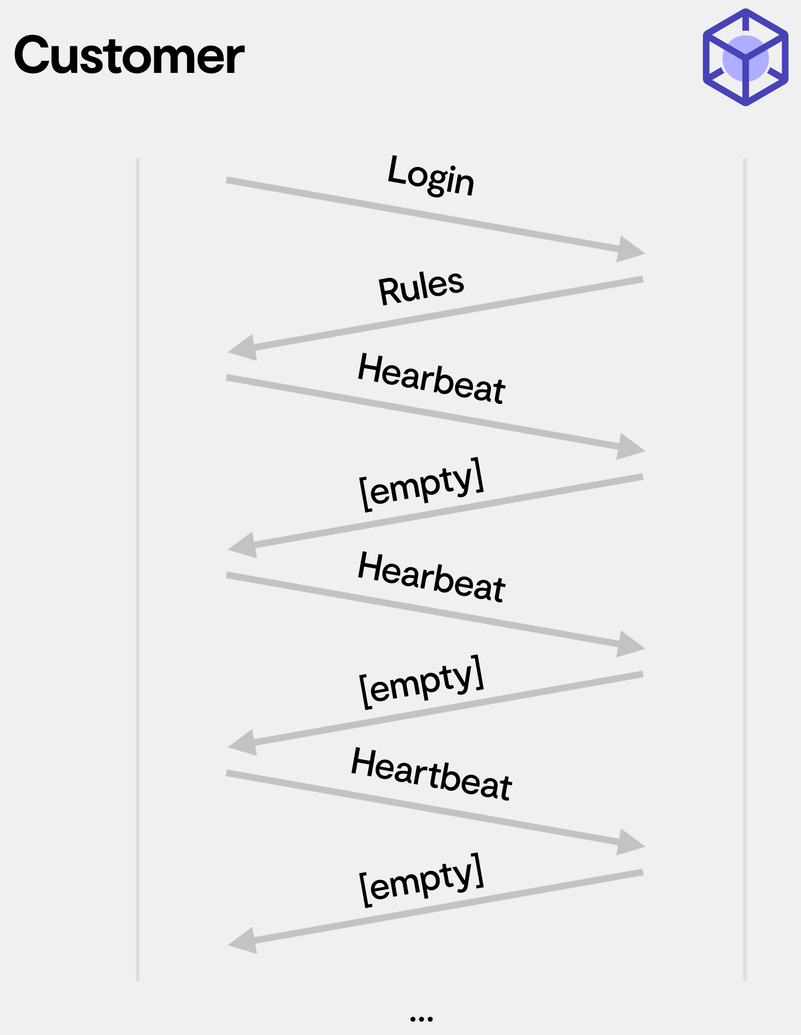
jb@sqreen.io
@jbaviat

What is Sqreen, how does it work?

Protects your app (HTTP)

Few big reads

Lots of small writes



Legal disclaimer

The information contained in this presentation is for general guidance on matters of interest only. The application and impact of laws can vary widely based on the specific facts involved. Given the changing nature of laws, rules and regulations, and the inherent hazards of electronic communication, there may be delays, omissions or inaccuracies in information contained in this presentation. Accordingly, the information on this site is provided with the understanding that the authors and publishers are not herein engaged in rendering legal, accounting, tax, or other professional advice and services. As such, it should not be used as a substitute for consultation with professional accounting, tax, legal or other competent advisers. Before making any decision or taking any action, you should consult a professional.

While we have made every attempt to ensure that the information contained in this site has been obtained from reliable sources, Keynote is not responsible for any errors or omissions, or for the results obtained from the use of this information. All information in this site is provided "as is", with no guarantee of completeness, accuracy, timeliness or of the results obtained from the use of this information, and without warranty of any kind, express or implied, including, but not limited to warranties of performance, merchantability and fitness for a particular purpose. In no event will Jb, its related partnerships or corporations, or the partners, agents or employees thereof be liable to you or anyone else for any decision made or action taken in reliance on the information in this Site or for any consequential, special or similar damages, even if advised of the possibility of such damages.

Certain links in this site connect to other websites maintained by third parties over whom Sqreen has no control. Sqreen makes no representations as to the accuracy or any other aspect of information contained in other websites.

sqreen

Legal disclaimer

The information contained in this presentation is for general guidance on matters of interest only. The application and impact of laws can vary widely based on the specific facts involved. Given the changing nature of laws, rules and regulations, and the inherent hazards of electronic communication, there may be delays, omissions or inaccuracies in information contained in this presentation. Accordingly, the information on this site is provided with the understanding that the authors and publishers are not herein engaged in rendering legal, accounting, tax, or other professional advice and services. As such, it should not be used as a substitute decision or taking any
While we h ble sources, Keynote is formation in this site is e use of this information, and without warranty of any kind, express or implied, including, but not limited to warranties of performance, merchantability and fitness for a particular purpose. In no event will Jb, its related partnerships or corporations, or the partners, agents or employees thereof be liable to you or anyone else for any decision made or action taken in reliance on the information in this Site or for any consequential, special or similar damages, even if advised of the possibility of such damages.
Certain links in this site connect to other websites maintained by third parties over whom Sqreen has no control. Sqreen makes no representations as to the accuracy or any other aspect of information contained in other websites.

PROD OUTAGES, YES BUT...

No impact on Sqreen customers production.

sqreen

0 RPM

Our current Kimsufi models

For all the following models, the setup fees are $13.99 plus tax

Model	CPU	Cores/ Threads	Freq.	RAM	Disk	Price/month		Quantity	
KS-6	Xeon 2xE5530	8c / 16t	2.4 GHz+	24 GB	2 TB	$43.00	120 sec	1	
KS-4C	Core™ i5-2300	4c / 4t	2.8 GHz+	16 GB	2 TB	$32.00	120 sec	1	
KS-3C	Core™ i3-2130/3240	2c / 4t	3.4 GHz+	8 GB	2 TB	$25.00	120 sec	1	
KS-2B	Atom™ N2800	2c / 4t	1.86 GHz+	4 GB	40 GB SSD	$14.00	120 sec	1	

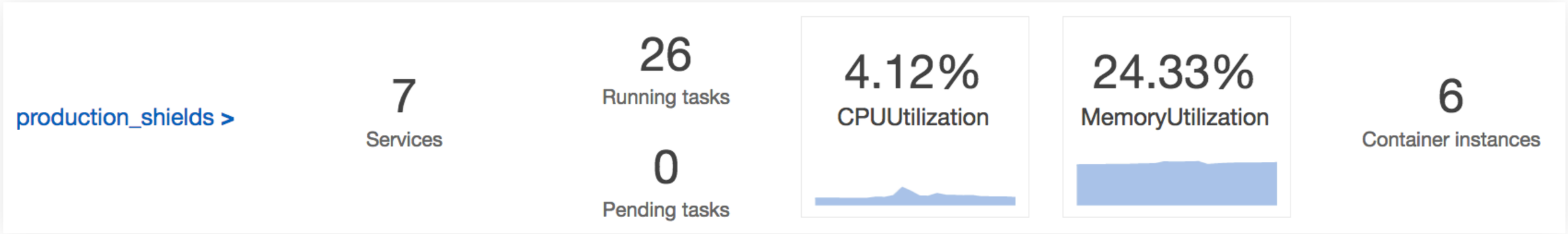
sqreen

10 RPM

10 RPM

AWS

- Free (startup in a co-working place)
- Docker capable (ECS)
- Security is great (*can be*)



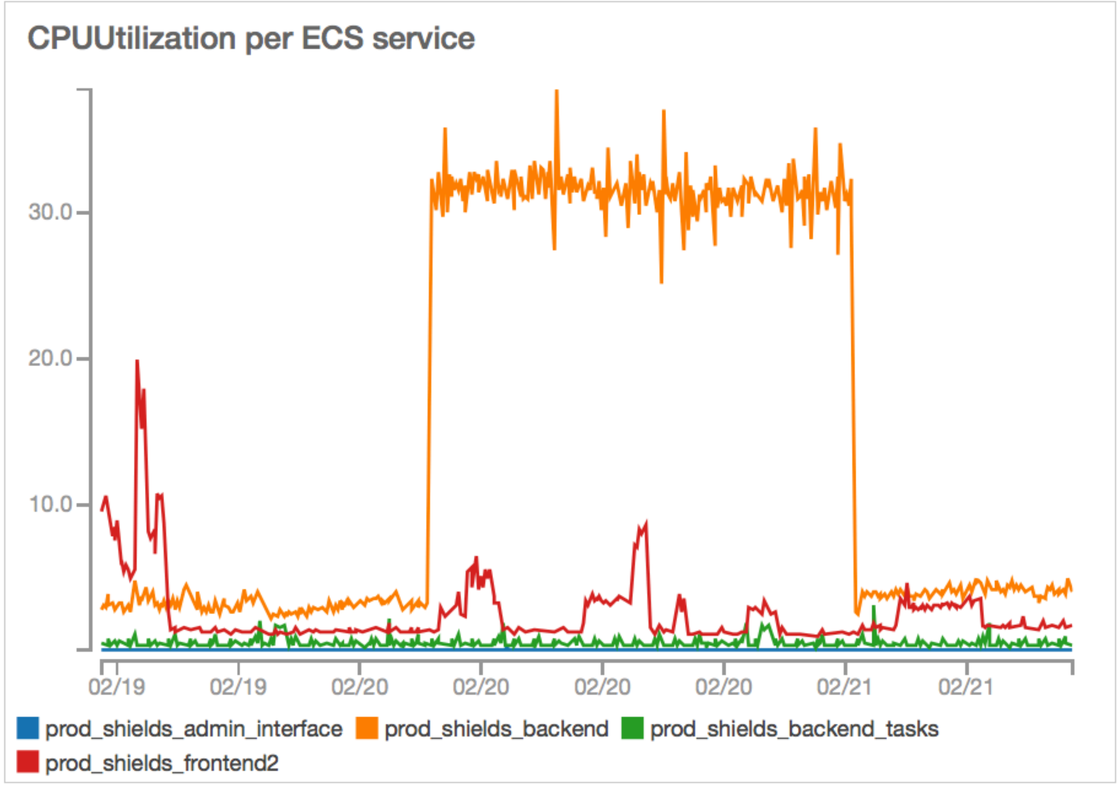
10 RPM

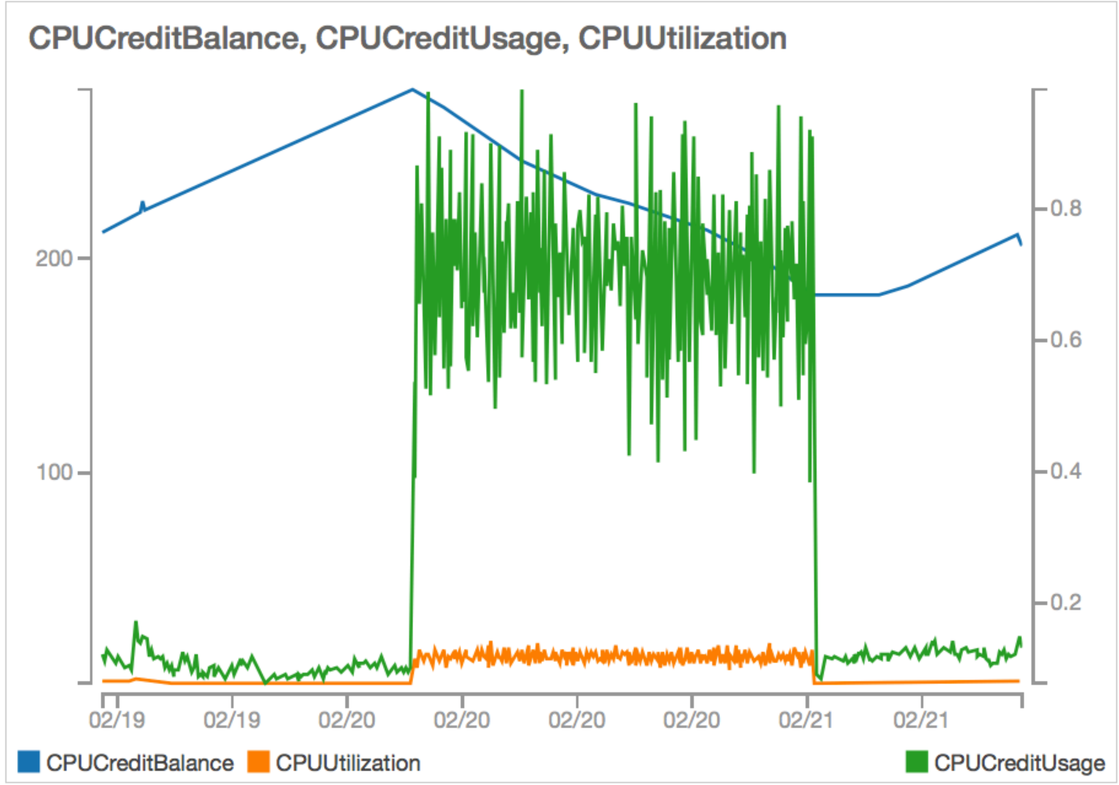
2015 = ECS early days

- Need 2 instances
- ELB need Docker to bind a static port
- You cannot bind the same port twice on a machine...
- No service interrupt on deploy: need 2 machines

10 RPM

t2 = burstable instances...





sqreen

100 RPM

100 RPM

First scaling issue

100 RPM

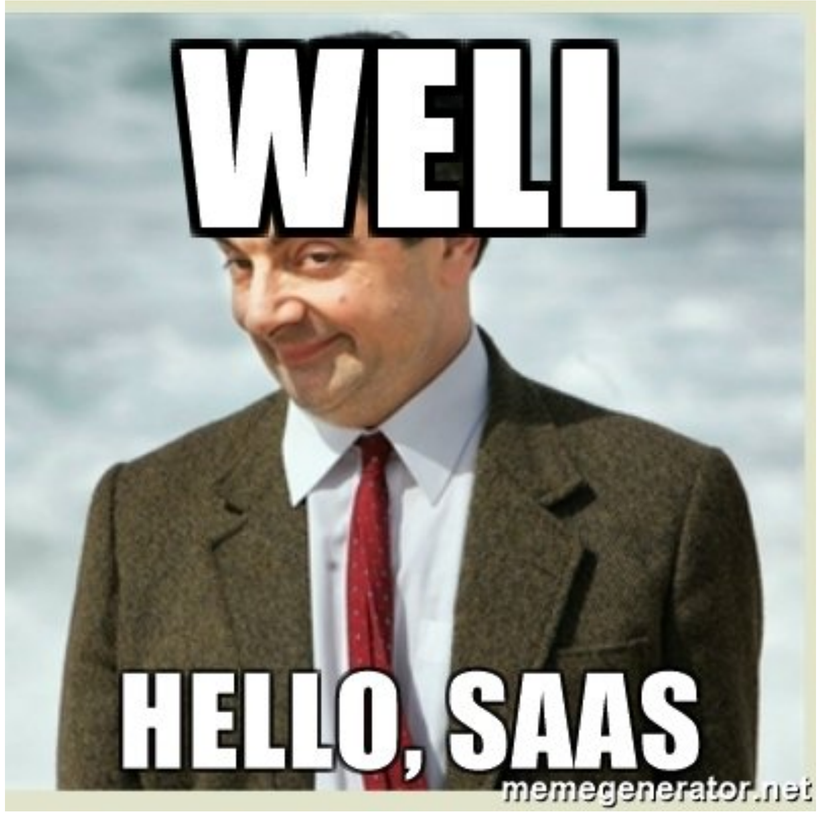
First scaling issue

Let’s boot more machines!

Keep focus on building the product

100 RPM

With > 1 service...



Read the logs? LOGGLY

Monitor the machines? New Relic®

Catch exceptions? SENTRY

100 RPM

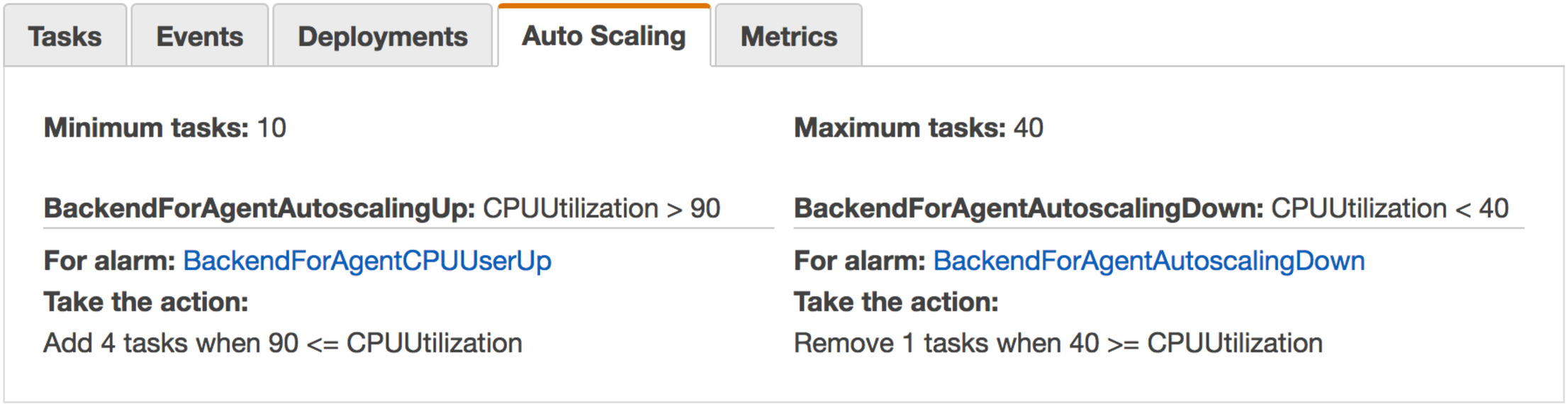
ALB (newer ELB) is released

- Removed 1 service per machine limitation
- Allows to build smaller services
- Allows per service auto scaling
- Enforce CPU limitations

100 RPM

Auto scaling

CPU bound: let's scale on CPU!



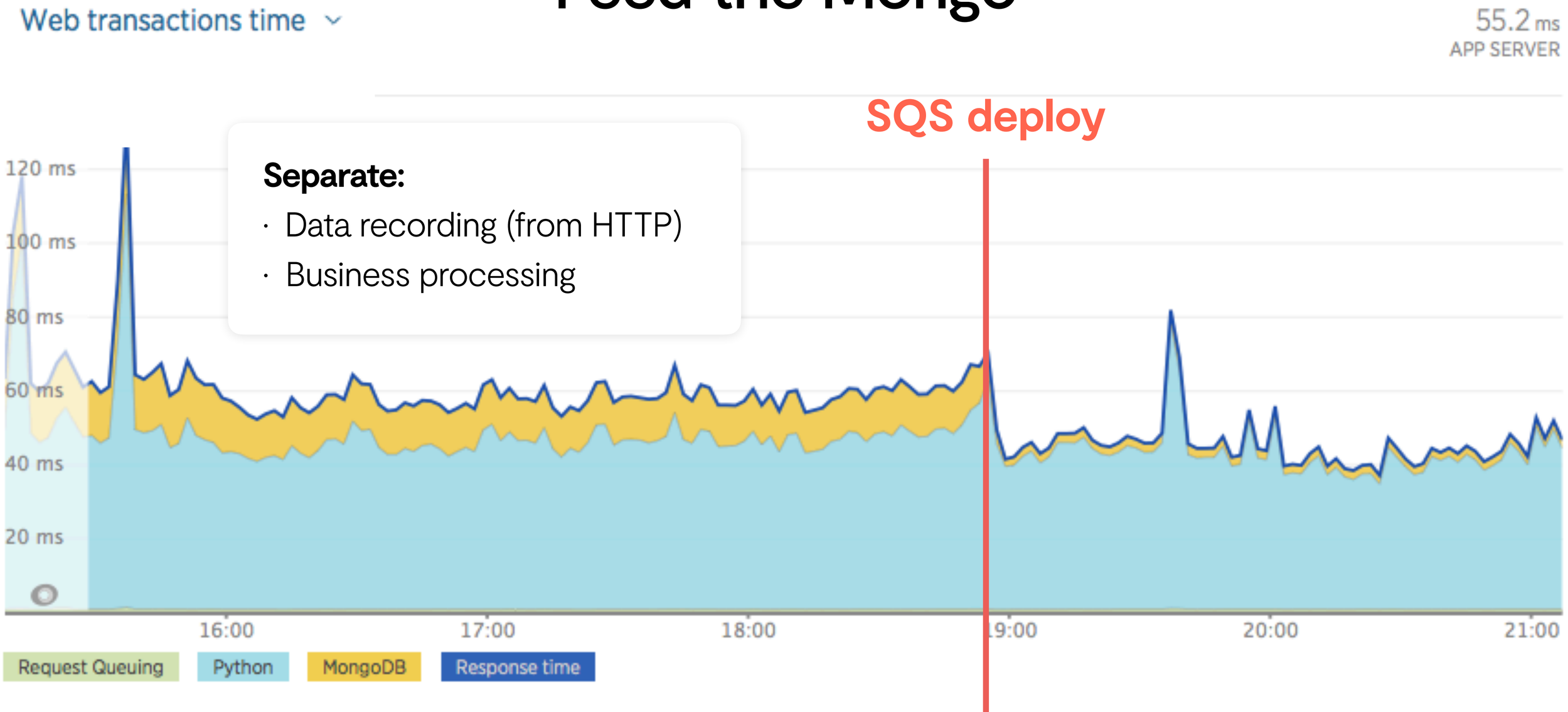
sqreen

1000 RPM

sqreen

1 000 RPM

Feed the Mongo



sqreen

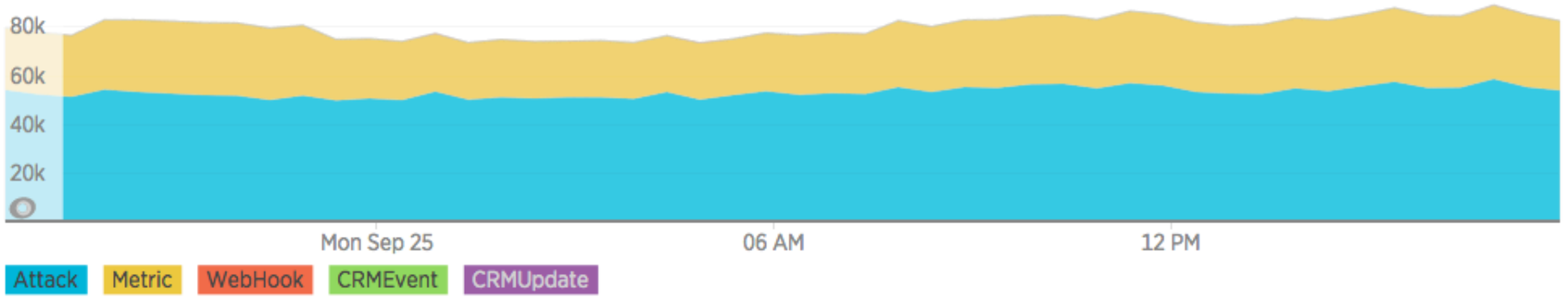
New Relic® INSIGHTS™

1 000 RPM

How to monitor SQS?

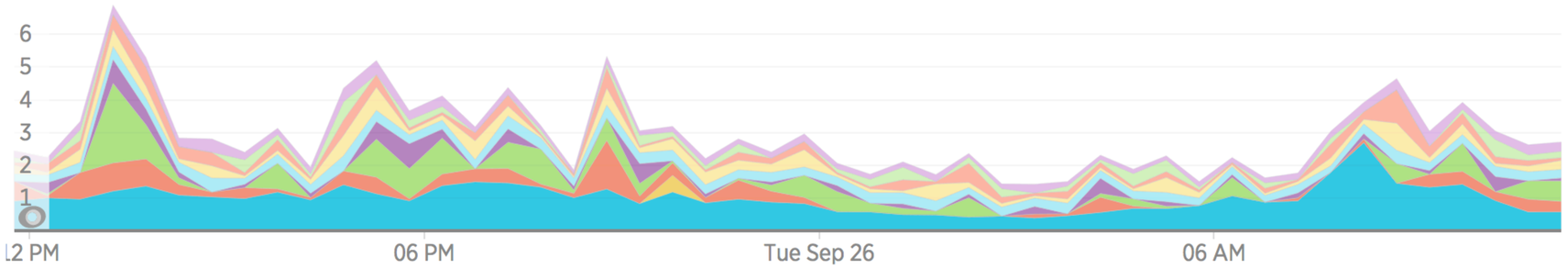
Digested number per kind

Since 1 day ago



Digestion time per customer

Since 1 day ago



sqreen

ALERT
Production Issue

sqreen

ALERT
Production Issue

- Login endpoint is taking too much time.
- The machines cannot take it anymore.
- RPM goes to 0.

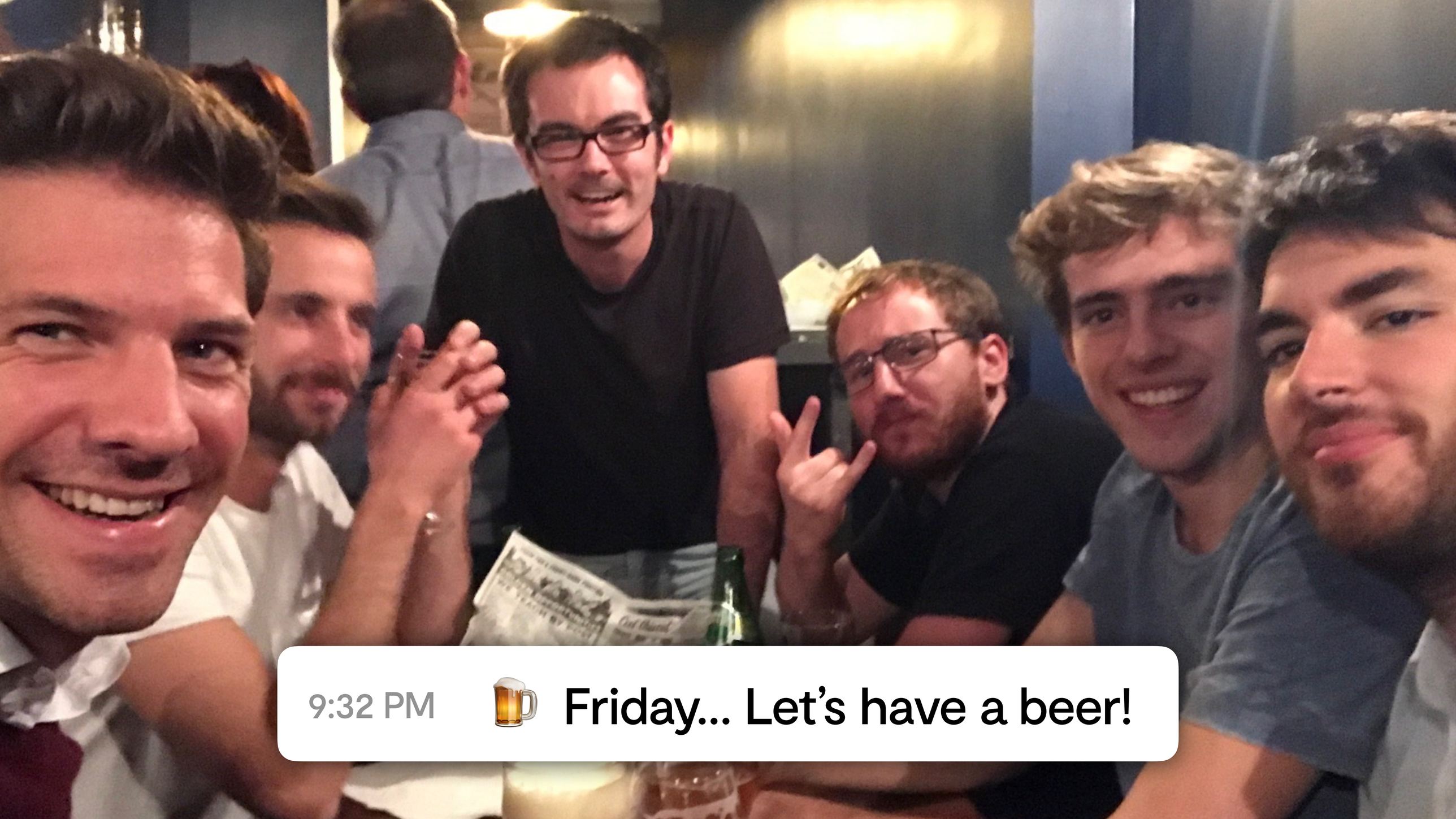
sqreen

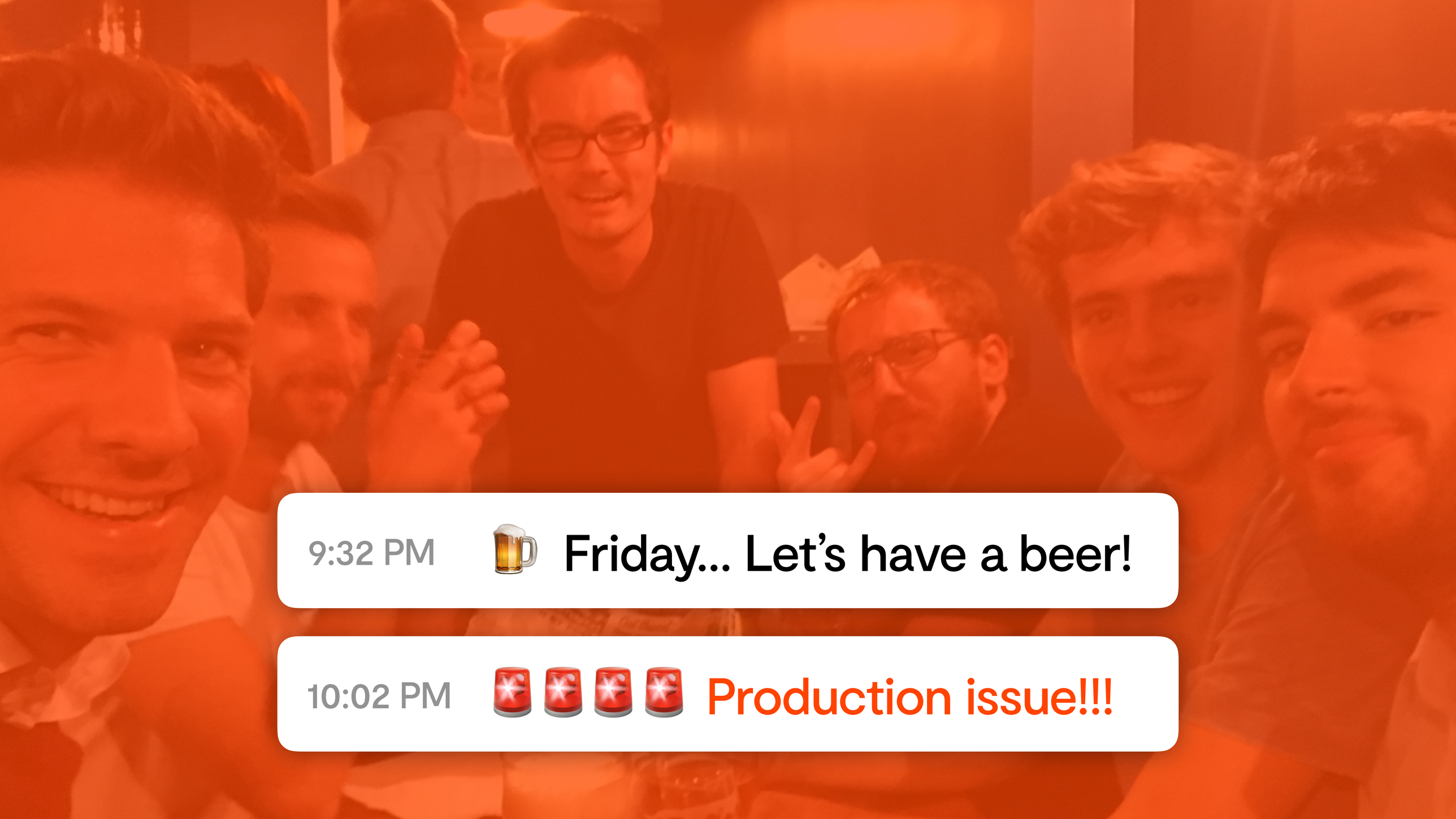
ALERT
Production Issue

- Login end

EMERGENCY FIX

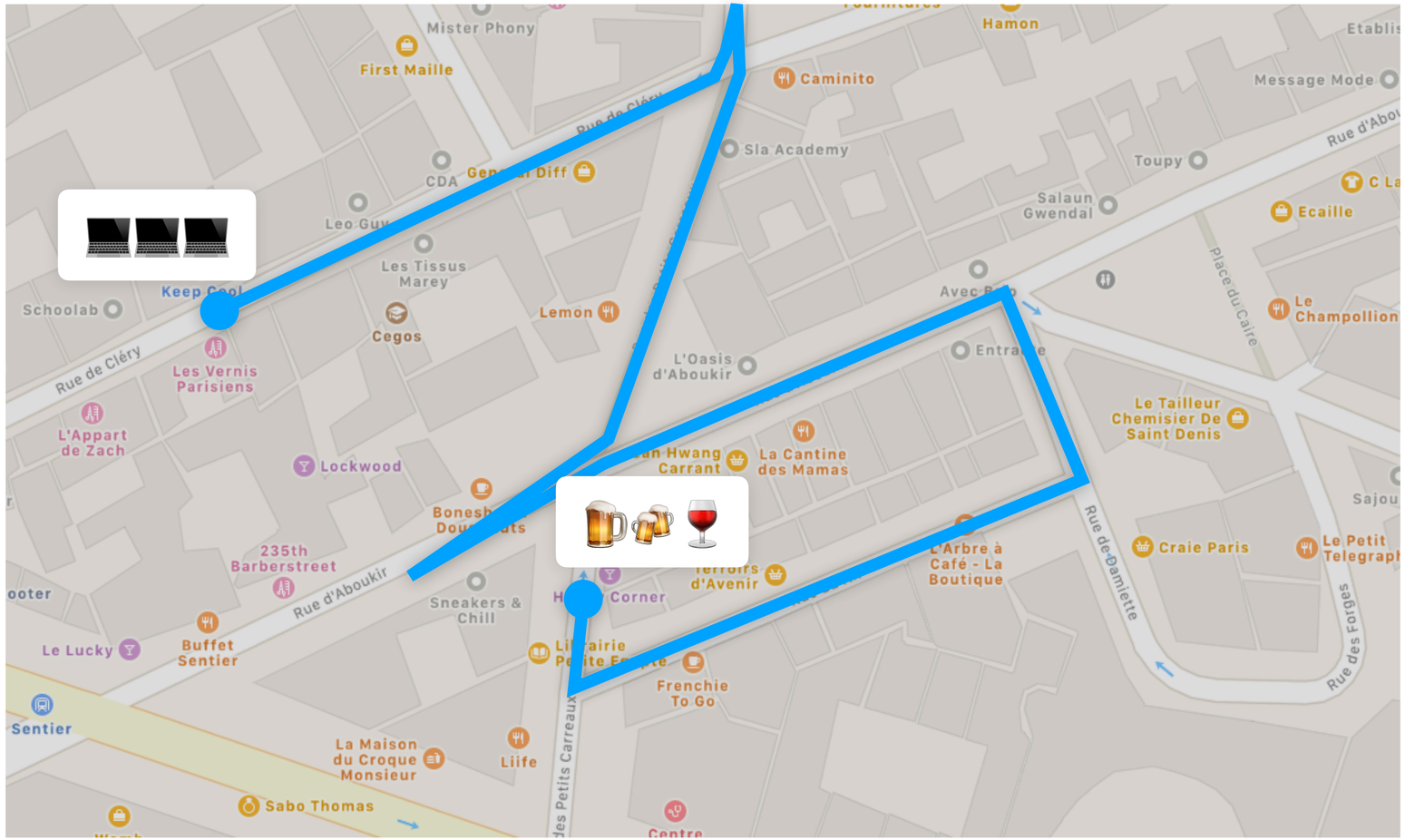
- Boot (way) more machines
- Use memcache to handle the login payload

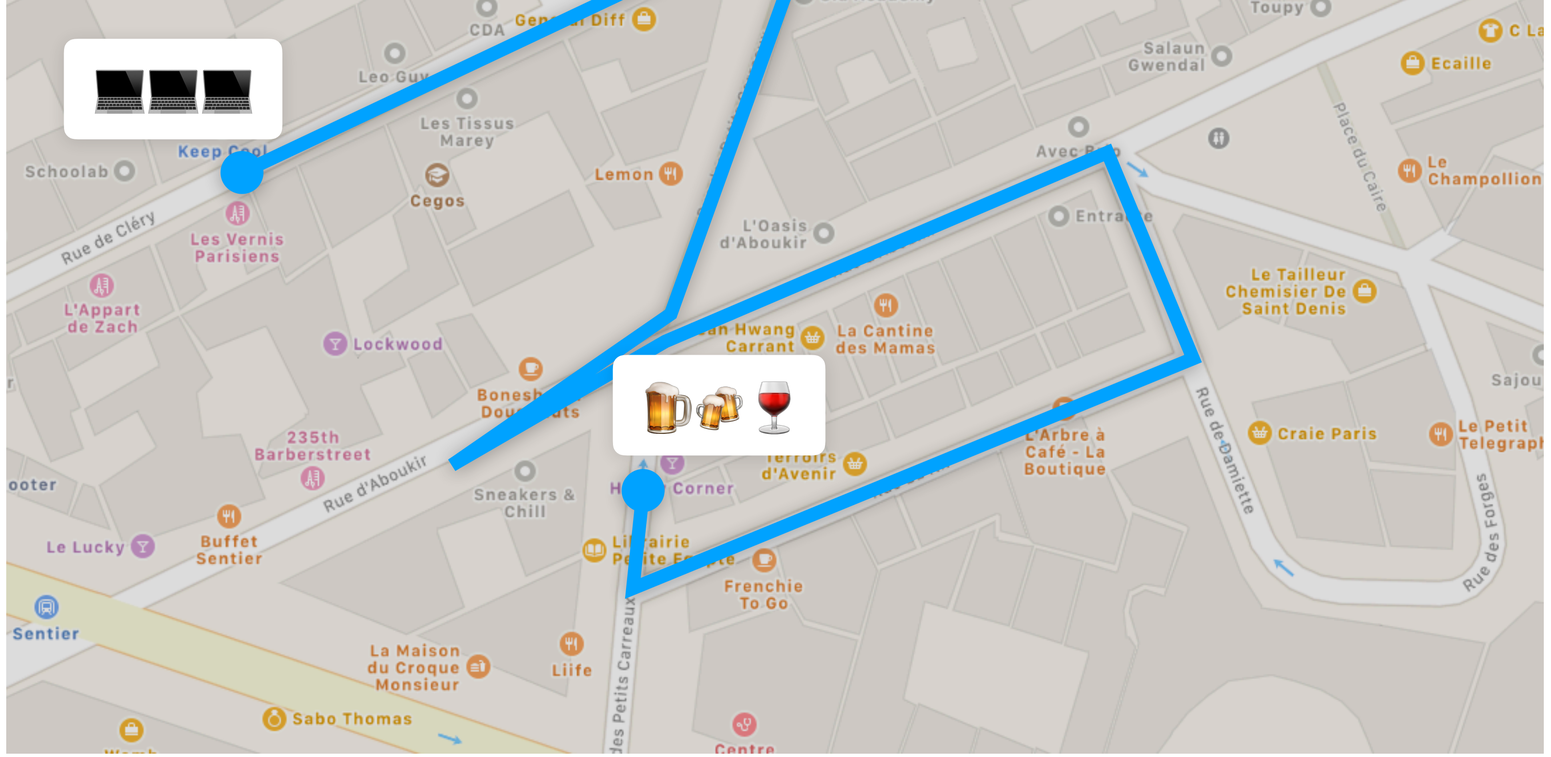

9:32 PM
Friday... Let’s have a beer!


9:32 PM
Friday... Let’s have a beer!
10:02 PM
Production issue!!!

9:32 PM 🍺 Friday... Let's have a beer!

10:02 PM 🚨🚨🚨🚨 Production issue!!!





10:25 PM

Big customer deploy

Friday evening

/login endpoint was (still) too slow

EMERGENCY FIX:

Boot (way) more machines

1 000 RPM

How do we fix this?

1

Pager Duty

Let’s get called!

2

Change agent/server protocol

Login was 4 requests
We made it 1 request

sqreen

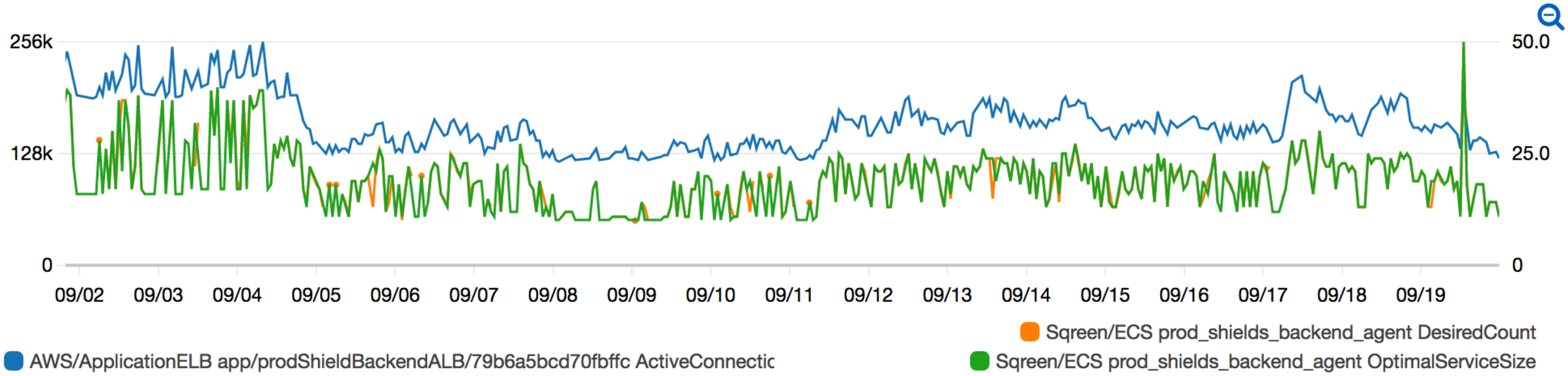
10 000 RPM

10 000 RPM

Auto scaling - Take 2

Need to scale ***faster***

Good metric: incoming requests



10 000 RPM

Auto scaling - Take 2

Better, but still ***too slow***...

We keep a “reserve”: services running all the time

Allow to handle spikes of new customers

sqreen

40 000 RPM

40 000 RPM

Now, we cannot fail anymore

Provisioned capacity.

Load testing:

- “Bees with machines guns” like
- With a realistic payload
- Simulate millions of servers using Sqreen
- Good tool to do so: Kubernetes

sqreen

60 000 RPM

60 000 RPM

Now we got SLAs

Queue + MongoDB... is not enough

—> Kinesis, DynamoDB

Better scaling

More resiliency to sudden loads

Lower operational costs

60 000 RPM

Next challenges

We're hiring!
sqreen.io/jobs

Smoother handling of specific customers

Reduce cost

Reduce latency

Move all our detection algorithms to streams

sqreen

Today

60 K

RPM

413 M

Attacks blocked last year

37 B

Requests protected last year

17 K

Attackers detected

We're hiring!
sqreen.io/jobs

Questions ?