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.
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.

PROD OUTAGES, YES BUT...

No impact on Sqreen customers production.
0 RPM
Our current Kimsufi models

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

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

AWS

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

2015 = ECS early days
10 RPM

t2 = burstable instances...
100 RPM
100 RPM

First scaling issue

OPTIMIZE

ALL THE THINGS!
Let’s boot more machines!

Keep focus on building the product
100 RPM
With > 1 service...

Read the logs?

Monitor the machines?

Catch exceptions?

LOGGLY

New Relic.

SENTRY
ALB (newer ELB) is released

- Removed 1 service per machine limitation
- Allows to build smaller services
- Allows per service auto scaling
- Enforce CPU limitations
CPU bound: let’s scale on CPU!

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Events</th>
<th>Deployments</th>
<th>Auto Scaling</th>
<th>Metrics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Minimum tasks:</strong> 10</td>
<td><strong>Maximum tasks:</strong> 40</td>
<td><strong>BackendForAgentAutoscalingUp:</strong> CPUUtilization &gt; 90</td>
<td><strong>For alarm:</strong> BackendForAgentCPUUserUp</td>
<td><strong>Take the action:</strong> Add 4 tasks when 90 &lt;= CPUUtilization</td>
</tr>
<tr>
<td><strong>BackendForAgentAutoscalingDown:</strong> CPUUtilization &lt; 40</td>
<td><strong>For alarm:</strong> BackendForAgentAutoscalingDown</td>
<td><strong>Take the action:</strong> Remove 1 tasks when 40 &gt;= CPUUtilization</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
1000 RPM
1 000 RPM

Feed the Mongo

Separate:
- Data recording (from HTTP)
- Business processing

SQS deploy
1000 RPM

How to monitor SQS?

Digested number per kind

Since 1 day ago

Digestion time per customer

Since 1 day ago
ALERT

Production Issue
**ALERT**

Production Issue

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

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

EMERGENCY FIX
- Boot (way) more machines
- Use memcache to handle the login payload
Friday... Let's have a beer!
Friday... Let's have a beer!

Production issue!!!
Friday... Let's have a beer!

Production issue!!!
10:25 PM  *Big* customer deploy

Friday evening

/\login endpoint was (still) too slow

**EMERGENCY FIX:**

Boot (way) more machines
1000 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
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
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
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
Next challenges

- Smoother handling of specific customers
- Reduce cost
- Reduce latency
- Move all our detection algorithms to streams

60,000 RPM

We're hiring!

sqreen.io/jobs
Today

- 60 K RPM
- 413 M Attacks blocked last year
- 37 B Requests protected last year
- 17 K Attackers detected
Questions?