

and zero security vulnerability shows a ZAP

create a locustfile.py in the root:

touch locustfile.py

import json

from locust import HttpUser, task, between

class AgnosticLambdaUser(HttpUser):

```

# Avarage wait time between tasks

wait_time = between(1, 2)

# Base URL for the Lambda function endpoint

@task

def test_lambda_endpoint(self):

    payload = {

        "test_key": "locust_test_v2",

        "source": "performance_validation"

    }

    headers = {

        "Content-Type": "application/json"

    }

    # Root path test

    with self.client.post("/", data=json.dumps(payload), headers=headers,
catch_response=True) as response:

        if response.status_code == 200:

            res_data = response.json()

            if res_data.get("status") == "success":

                response.success()

            else:

                response.failure(f"Unexpected status logic: {res_data}")

        else:

            response.failure(f"HTTP Error: {response.status_code}")

# Test for security headers on non-existent route

@task(1)

def test_security_route_protection(self):

    with self.client.get("/robots.txt", catch_response=True) as response:

        if response.status_code == 404:

            if "Strict-Transport-Security" in response.headers:

```

```
response.success()
else:
    response.failure("Security headers missing from 404 response")
else:
    response.failure(f"Route protection failed: {response.status_code}")
```

```
git add locustfile.py
git commit -m "feat: add locustfile.py"
git push
```

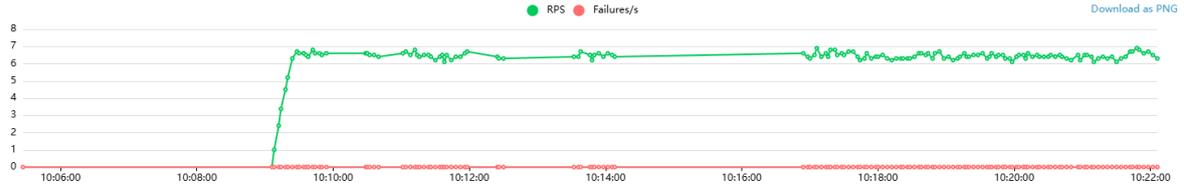
than terminal:
locust -f locustfile.py
open a browser: <http://localhost:8089>
and start test

Start new load test

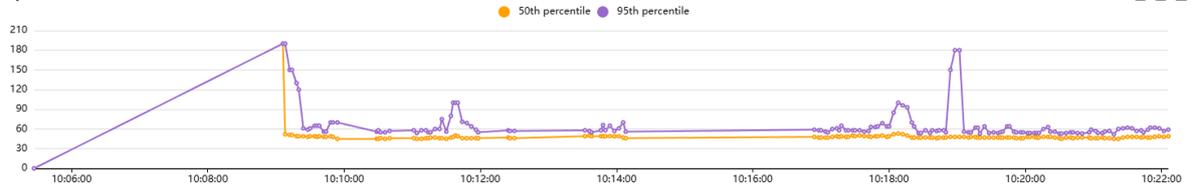
Number of users (peak concurrency) *

Ramp up (users started/second) *

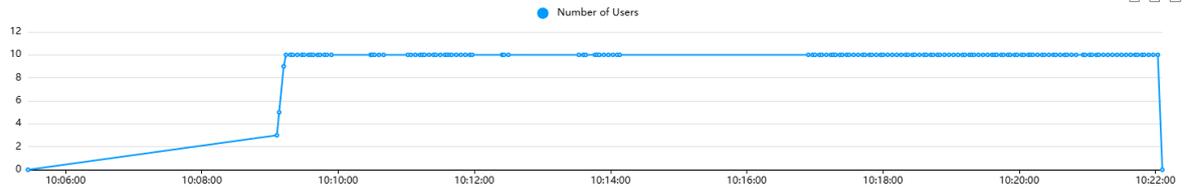
Total Requests per Second



Response Times (ms)



Number of Users



ok than a new test:

Start new load test

Number of users (peak concurrency) *

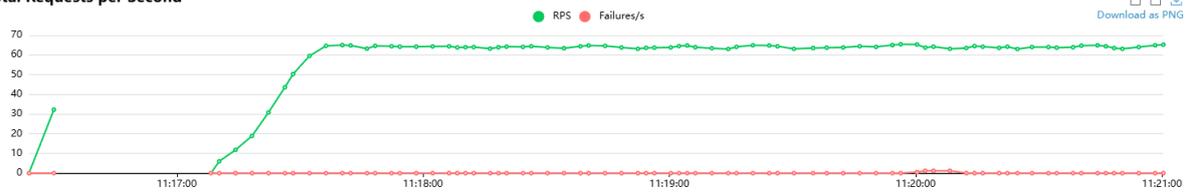
100

Ramp up (users started/second) *

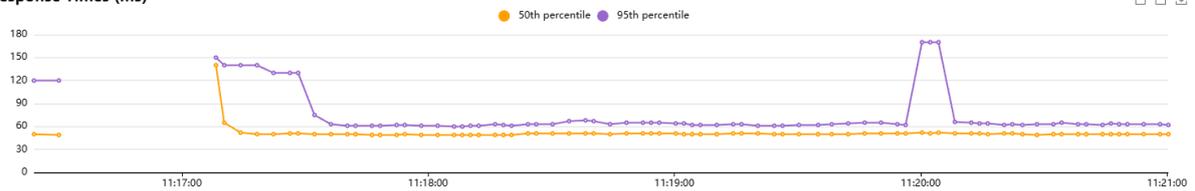
5

and

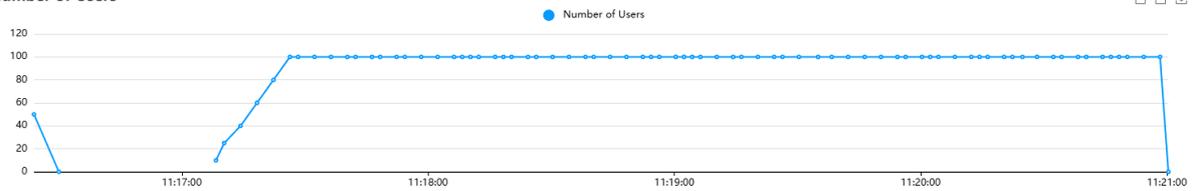
Total Requests per Second



Response Times (ms)



Number of Users



and one more new test: 200 users 5 ramp up and i identify between 150 and 180 my system starts failures

LOCUST Host: `https://ot3mexvbkmp55xdnufzgsqp6q0gwjuc.lambda...` Status: SPAWNING Users: 180 RPS: 94.3 Failures: 4% EDIT STOP RESET ⚙️

STATISTICS CHARTS **FAILURES** EXCEPTIONS CURRENT RATIO DOWNLOAD DATA LOGS LOCUST CLOUD

# Failures	Method	Name	Message
43	POST	//	CatchResponseError('HTTP Error: 429')
36	GET	//robots.txt	CatchResponseError('Route protection failed: 429')

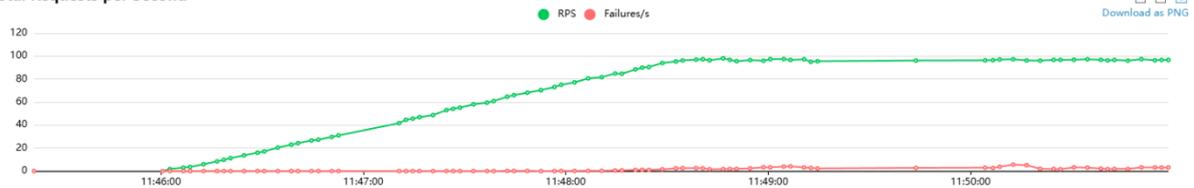
so check this 150 with slow ramp up:

Start new load test

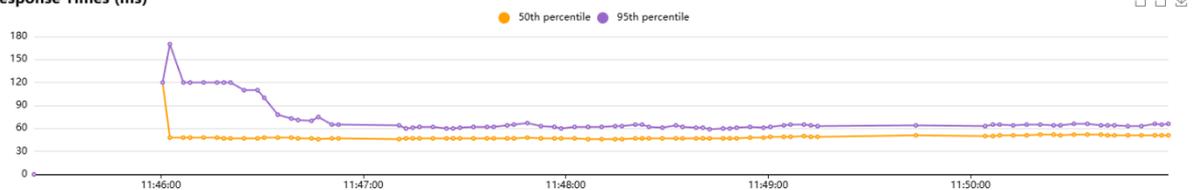
Number of users (peak concurrency) *

Ramp up (users started/second) *

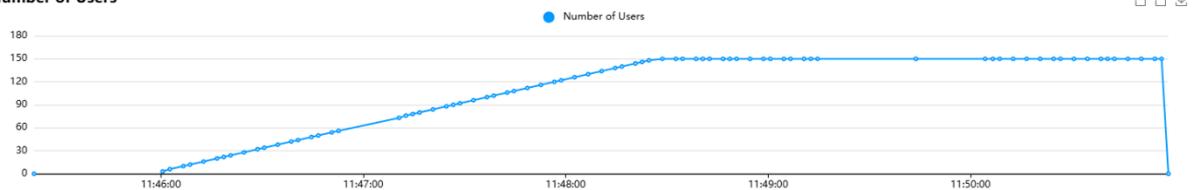
Total Requests per Second



Response Times (ms)



Number of Users



ctrl+c to stop a locust and get terminal to get a nice html file to your test:

```
mkdir locust_reports
```

```
locust -f locustfile.py --headless -u 150 -r 1 --run-time 6m --host https://<YOUR_LAMBDA_URL> --html locust_reports/v2_performance_report_150.html
```

time to level up: (the final test but it can cost you a money so beware!)

ctrl+c to stop a last locust test a terminal and:

```
locust -f locustfile.py --headless -u 1000 -r 10 --run-time 3m --host https://<YOUR_LAMBDA_URL> --html locust_reports/v2_performance_report.html
```

Final conclusion & engineering summary

The transition from v1 (Elastic Beanstalk/django) to v2 (agnostic serverless Lambda) has been technically validated with the following results:

1. Architectural decoupling: The implementation of the hexagonal/agnostic core ensures that the business logic is independent of the cloud provider's entry points.
2. Performance leap: Median response times were reduced from ~120ms (v1) to 44ms (v2), representing a 2.5x speed increase in request processing.
3. Hardened security: The 0-alert OWASP ZAP status and the Locust-validated HSTS/Security headers ensure a production-ready security posture.
4. Defined limits: Load testing identified the current infrastructure ceiling at approx. 150 concurrent users under standard AWS regional quotas.

Scalability Note: Testing threshold at 150 concurrent users limit is a default AWS Service Quota (Soft Limit-1000). The system scales horizontally; this threshold can be increased via an AWS Quota Request.

Status: Technical Validation Complete. The core logic and infrastructure performance are validated for production-level traffic.

At this Proof of Concept (PoC) level, implementing automated rollbacks and self-healing mechanisms would have introduced unnecessary configuration complexity. Consequently, measuring and optimizing professional DORA metrics—specifically automated MTTR—is intentionally deferred to the v3 (Infrastructure as Code) phase, where full operational automation will be established.