Locust Load Testing

This activity introduces students to the concept of application scaling through load testing using Locust, a Python-based open-source tool. Students will simulate high user traffic, measure system performance, identify bottlenecks, and evaluate how system design affects scalability.

Setup Requirements

- Install Locust

Unset

```
pip install locust
```

- Create a `locustfile.py` and add the following code:

Python

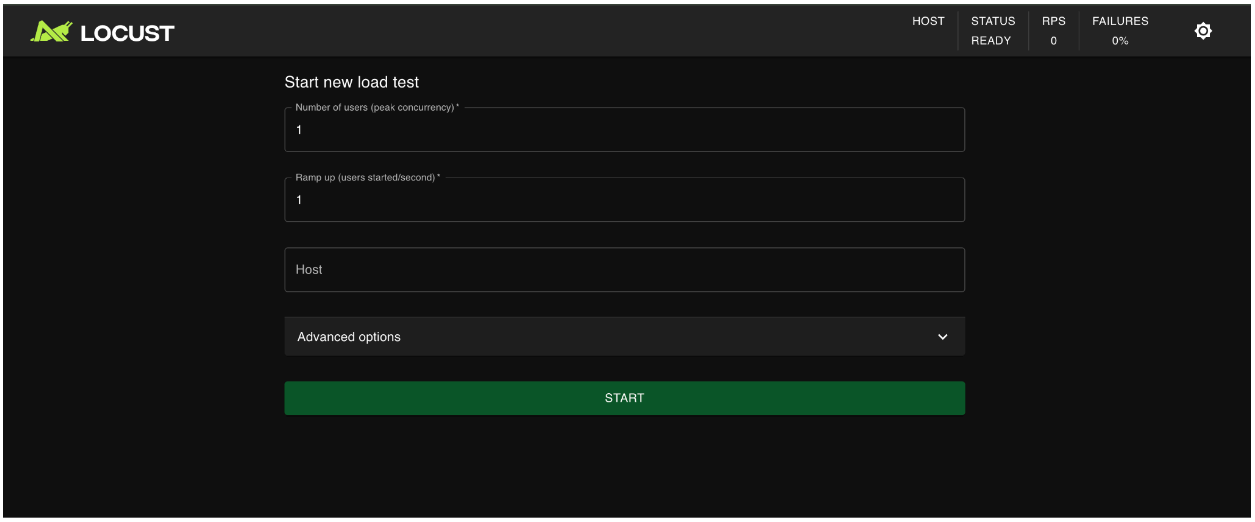
```python
from locust import HttpUser, task, constant

class UserBehavior(HttpUser):
    wait_time = constant(0)

    # This is a sample API call, change it according to your endpoint and method. Refer to Locust documentation for more details.
    @task
    def get_items(self):
        self.client.get("/items")
```

- **Run command:**

Unset

```
locust
```

- Visit the Locust web UI at: http://localhost:8089. You will see a screen like this:



Start new load test

Number of users (peak concurrency) *
1

Ramp up (users started/second) *
1

Host

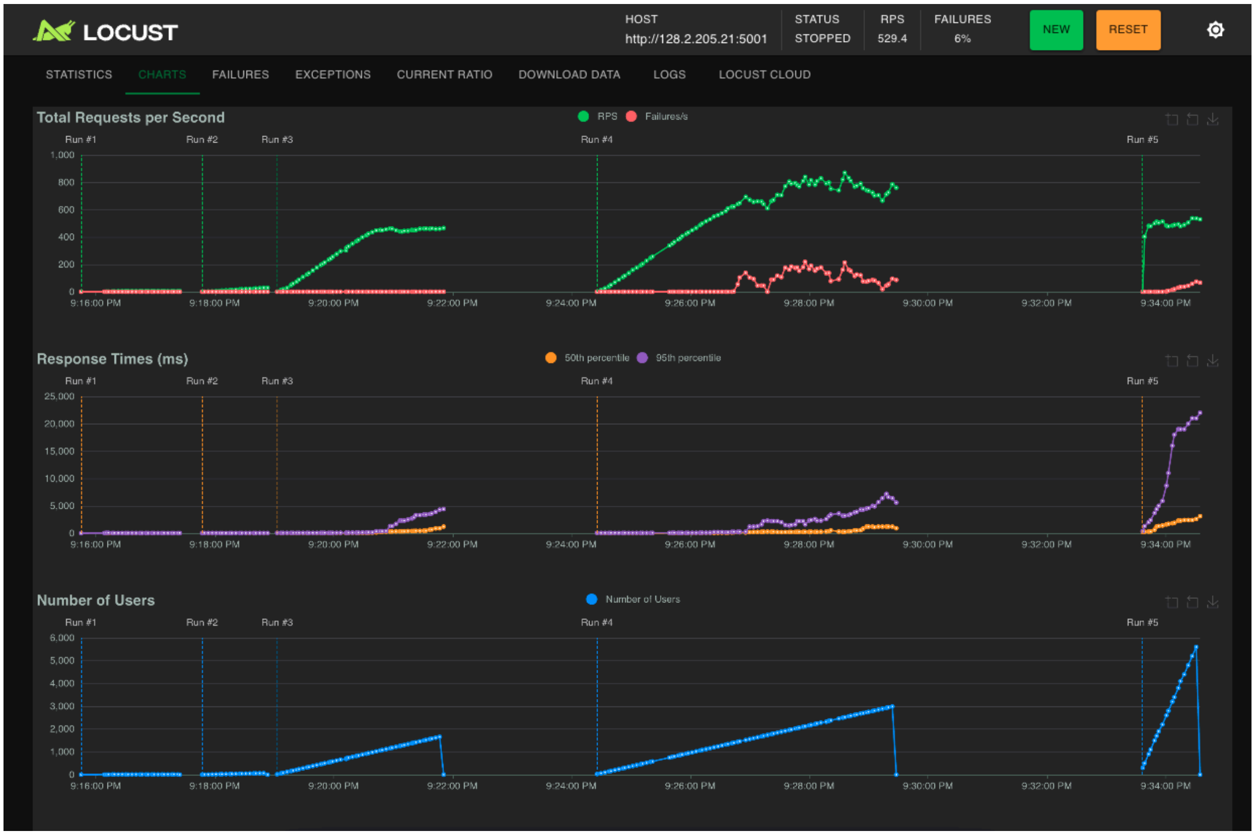
Configure and Run the Test

- Enter the **number of users to simulate** (e.g., 10000): This represents the maximum number of users that will be simulated concurrently to send the request
- Set a **spawn rate** (users per second): Number of new users that will be spawned every second.
- Set the **host** (e.g., http://17423-teamXX.s3d.cmu.edu/): The base API url where the endpoints are hosted.

Monitor and Analyze

Observe real-time metrics:

- Requests per second (RPS) - The number of requests the server is able to handle.
- Response time (average, median, p90) - Time it takes to respond to a service.
- Failures/s and error rates - Number of requests failing.



HOST http://128.2.205.21:5001
STATUS STOPPED
RPS 529.4
FAILURES 6%

Observe and Identify

- The threshold for RPS: What is the threshold for the number of requests your server can handle?
- p50 and p95 for response time
    - p50: What is the response time for the 50th percentile request.
    - P95: What is the response time for the 95th percentile request.
    - 95th percentile request means if we sort all the response time and there are 100 requests, the request with 5th highest response time is p95.
- Types of failures (In “Failures” tab on locust)

Improve

- How can we increase the RPS?
    - Try to identify the bottlenecks in your system and discuss how you can improve it.
- How can we reduce the response time?
    - Try to find inefficiencies in your system and discuss how you can make it efficient and reduce the response time per request.