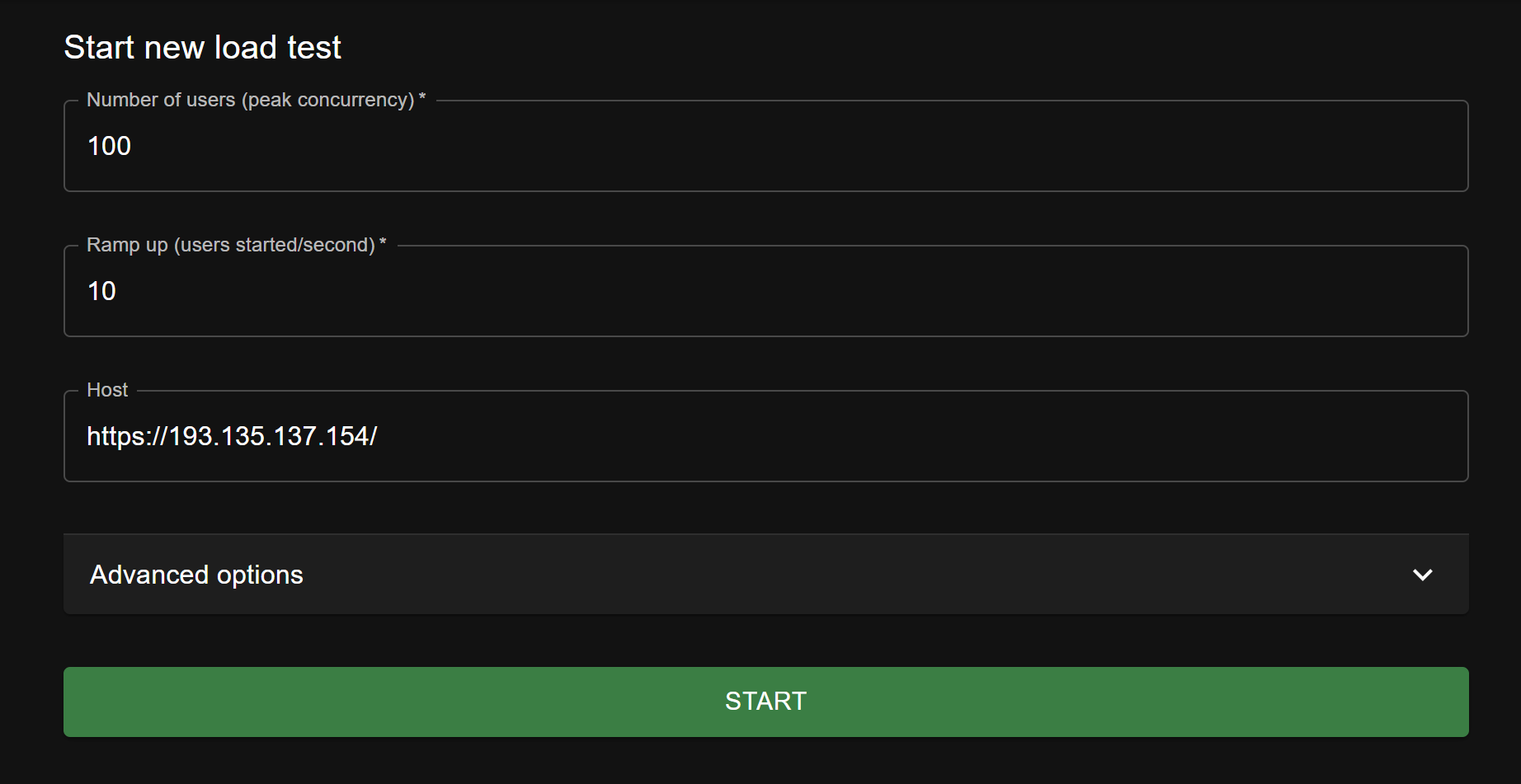
1. **Тест 1**

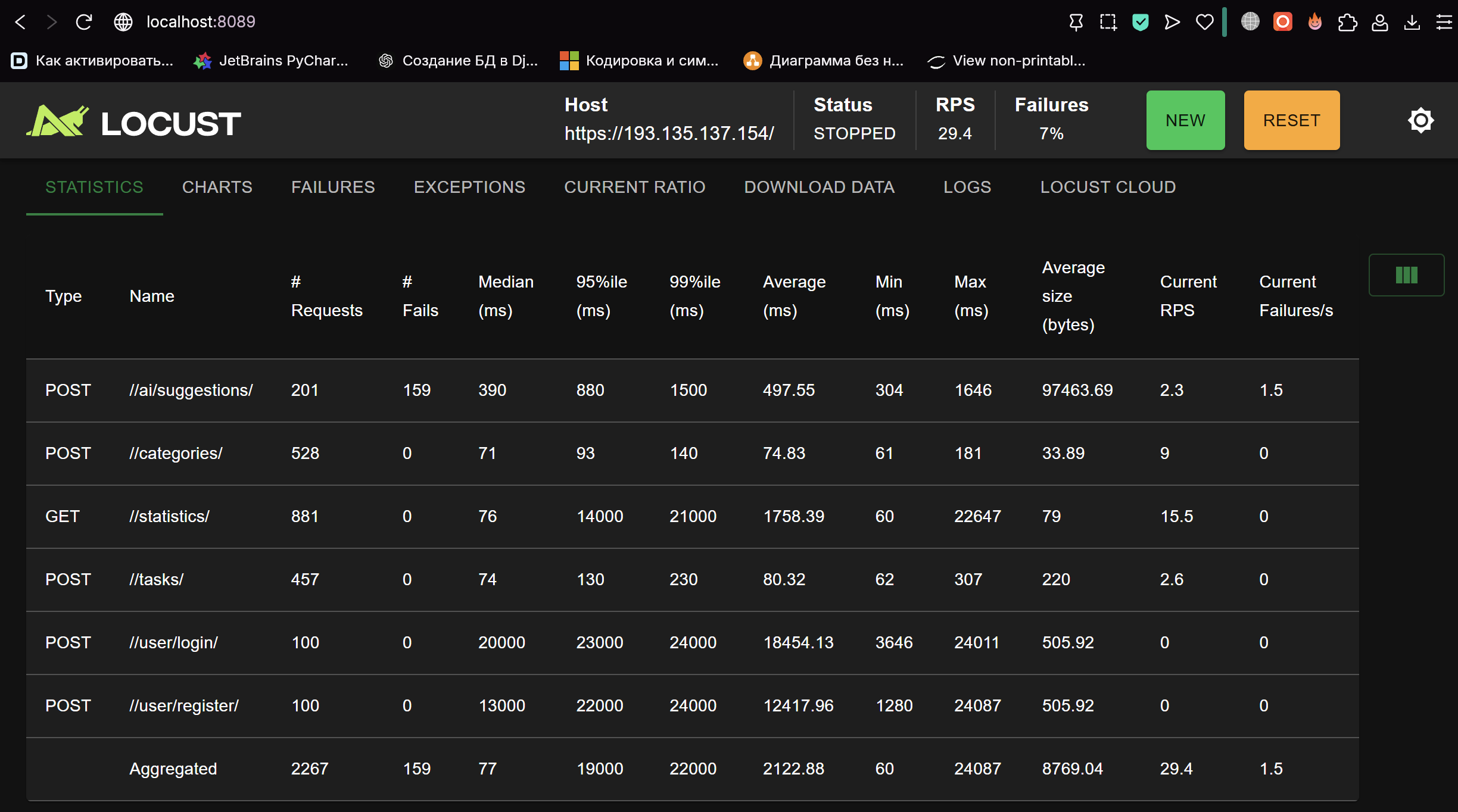
****

**Ключевые моменты:**

1. **Общая производительность**:
   1. **Total Requests**: 2267.
   2. **Total Fails**: 159 (7% от всех запросов).
   3. **Median**: 77 мс (хорошо для большинства эндпоинтов).
   4. **99%ile**: 2200 мс (пик задержек).
   5. **Average**: 2122.88 мс (сильно искажено медленными эндпоинтами).
   6. **Current RPS**: 29.4 (при 100 пользователях это умеренная нагрузка).
2. **Успешные эндпоинты**:
   1. /categories/: 528 запросов, 0 ошибок, медиана 71 мс, 99%ile 140 мс.
   2. /tasks/: 457 запросов, 0 ошибок, медиана 74 мс, 99%ile 230 мс.
   3. /statistics/: 881 запросов, 0 ошибок, медиана 76 мс, но 99%ile 2100 мс и максимум 22647 мс (выбросы).
   4. /user/login/: 100 запросов, 0 ошибок, но медиана 20000 мс (20 секунд!).
   5. /user/register/: 100 запросов, 0 ошибок, медиана 13000 мс (13 секунд!).

Медиана 20000 мс (20 сек) для /user/login/ и 13000 мс (13 сек) для /user/register/ — это аномально долго.

1. **Проблемный эндпоинт**:
   1. /ai/suggestions/: 201 запросов, 159 ошибок (79%!), медиана 390 мс, 99%ile 1500 мс, максимум 1646 мс.
   2. Это основной источник сбоев (1.5 Failures/s)
2. **Нагрузка**:
   1. **RPS 29.4** при 100 пользователях — это низкая нагрузка относительно лимита Nginx (1024).
   2. С Ramp up = 10 users/second, разгон занял 10 секунд (100 / 10)



**Однопоточная обработка GigaChat**:

1. Бесплатный тариф GigaChat позволяет обрабатывать только один запрос одновременно. Если приходит второй запрос, пока первый выполняется, он либо откладывается, либо завершается с ошибкой (например, таймаутом), что приводит к 500 ошибке на стороне сервера.
2. В тесте с 100 пользователями и 2.3 RPS на /ai/suggestions/ (201 запроса, 159 ошибок) нагрузка превысила возможности однопоточного тарифа, вызвав перегрузку.

**Задержки**:

Медиана 390 мс и максимум 1646 мс для /ai/suggestions/ показывают, что каждый запрос занимает заметное время. При однопоточной модели новые запросы накапливаются, и сервер не успевает их обработать, что приводит к ошибкам.

Но это можно исправить если:

1. обучить свою llm и развернуть на нескольких мощных машинах
2. платить за GigaChat

Для первого варианта необходимы капитальные затраты на аренду вычислительных мощностей

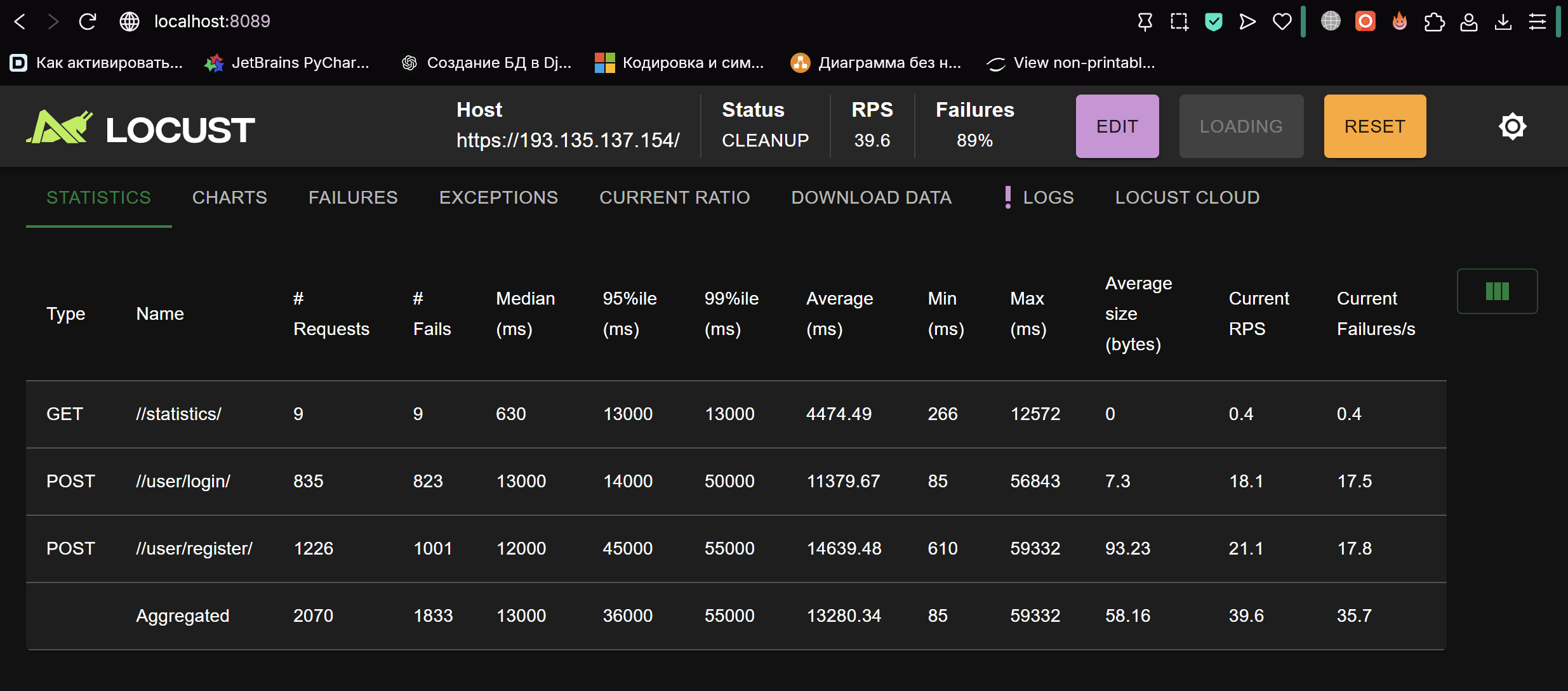
Второй подход предполагает операционные расходы на подписку

Оба варианта решаются только после релиза или после получения денежных средств, но на первых парах данных мощностей хватит.

### Тест 2

### Характеристика теста (500 пользователей, Ramp up = 20 users/second)

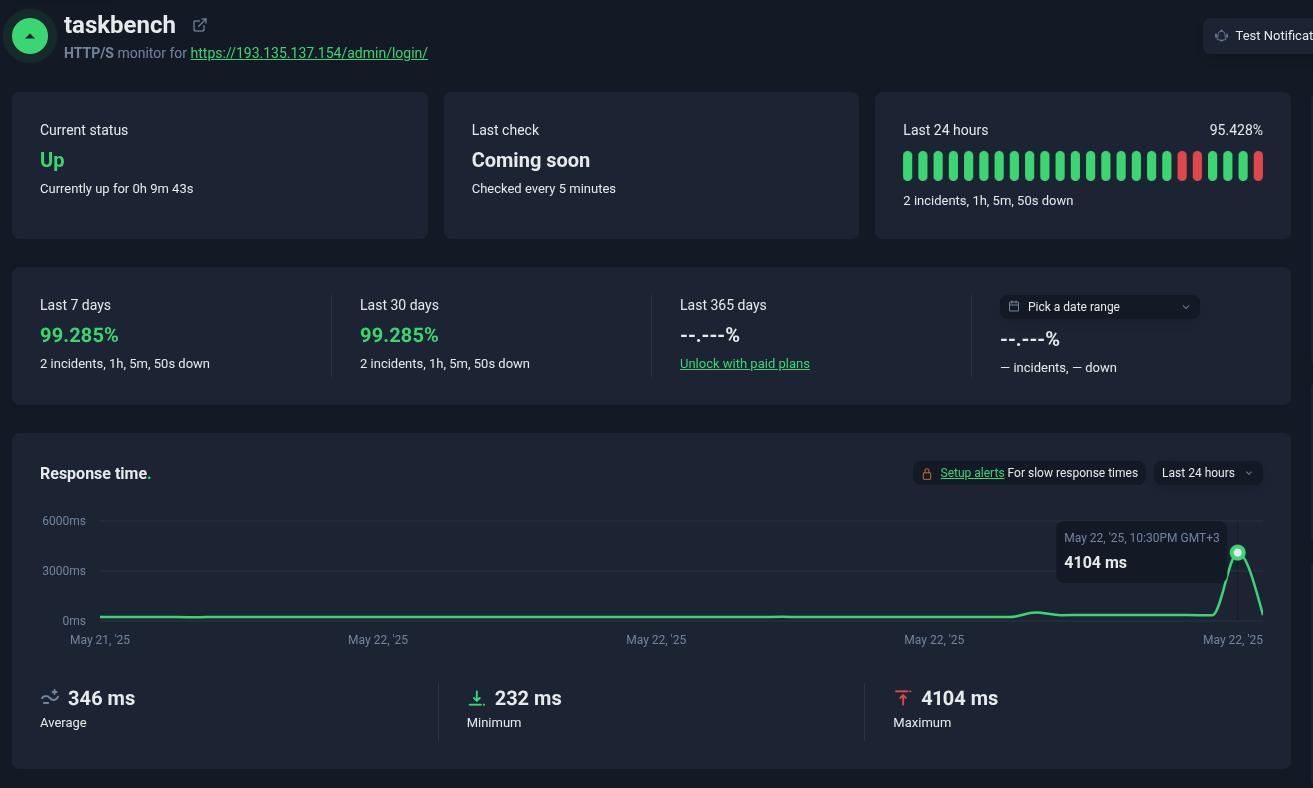
1. **Общая производительность**:
   1. **Total Requests**: 2294.
   2. **Total Fails**: 524 (22.8% ошибок — значительный процент).
   3. **Median**: 6000 мс (6 секунд) — высокая задержка.
   4. **95%ile**: 11400 мс (11.4 секунды).
   5. **99%ile**: 11500 мс (11.5 секунды).
   6. **Average**: 5813.59 мс (5.8 секунды).
   7. **Current RPS**: 19.3 (умеренная нагрузка).
   8. **Current Failures/s**: 0.8.
2. **Детали по эндпоинтам**:
   1. **/categories/**:
      * 1. 27 запросов, 4 ошибки (14.8%), медиана 3800 мс, максимум 7421 мс.
        2. Задержки высокие, есть сбои.
   2. **/statistics/**:
      * 1. 753 запроса, 185 ошибок (24.6%), медиана 5000 мс, максимум 11524 мс.
        2. Значительные задержки и сбои.
   3. **/user/login/**:
      * 1. 739 запросов, 292 ошибки (39.5%), медиана 5700 мс, максимум 11595 мс.
        2. Самый проблемный эндпоинт с высоким процентом ошибок.
   4. **/user/register/**:
      * 1. 775 запросов, 43 ошибки (5.5%), медиана 7500 мс, максимум 11604 мс.
        2. Задержки высокие, но меньше ошибок, чем у /user/login/.
3. **Нагрузка**:
   1. **RPS 19.3** при 500 пользователях — это ниже лимита Nginx (1024 подключения).
   2. С Ramp up = 20 users/second, 500 пользователей добавились за 25 секунд (500 / 20). Общая длительность теста, вероятно, 5 минут (--run-time=5m), что дало время для накопления запросов.

**Тест 3**  
тест с 1000 пользователями и Ramp up = 100 users/second  
  


#### Ключевые моменты

1. **Общая производительность**:
   1. **Total Requests**: 2070.
   2. **Total Fails**: 1833 (88.5% от всех запросов!).
   3. **Median**: 13000 мс (13 секунд) — очень высокая задержка.
   4. **99%ile**: 55000 мс (55 секунд) — критично долго.
   5. **Average**: 13280.34 мс (13.3 секунды).
   6. **Current RPS**: 39.6.
   7. **Current Failures/s**: 35.7 — огромное количество ошибок в секунду.
2. **Детали по эндпоинтам**:
   1. **/statistics/**:
      * 1. 9 запросов, 9 ошибок (100%), медиана 630 мс, но 99%ile 13000 мс и максимум 12572 мс.
        2. Вероятно, сервер не успевал отвечать.
   2. **/user/login/**:
      * 1. 835 запросов, 823 ошибки (98.6%), медиана 13000 мс, максимум 56843 мс.
        2. Задержки огромные, большинство запросов провалилось.
   3. **/user/register/**:
      * 1. 1226 запросов, 1001 ошибка (81.6%), медиана 12000 мс, максимум 59332 мс.
        2. Аналогично, сервер не справился.

**Вывод:**



Пик в 4104 мс был зафиксирован в 10:30 PM GMT+3 (22 мая 2025), совпадая с нагрузочным тестом (500 пользователей). Это, вероятно, вызвано перегрузкой сервера, особенно на /user/login/ и /statistics/ После пика время отклика стабилизировалось (среднее 346 мс).