LG U+ Why Not SW Camp 7기

CookUs

성능 평가 결과 보고서

TEAM 건전지

홍가연, 박민상, 김지영, 신윤서

목 차

프로젝트명**: Cookus –** 냉장고 기반 맞춤형 레시피 추천 서비스

I. 테스트 환경

항목	내용
서버 환경	AWS EC2 **t3.micro** (2 vCPU / 1GB RAM) – Docker 기반 배포, Nginx + Uvicorn
운영체제	Amazon Linux 2023 kernel-6.1 AMI
백엔드 프레임워크	FastAPI (Python 3.10)
DB	MariaDB 10.6 (AWS RDS), S3
네트워크	1Gbps
테스트 도구	Apache Benchmark(ab), htop, FastAPI built-in logging

II. 테스트 목적

본 평가는 Cookus 서비스의 성능 안정성**,** 응답 속도**,** 동시 사용자 처리 능력을 확인하고,

특히 로그인 · 레시피 추천 · 커뮤니티 게시글 등록 등 핵심 기능에서의 병목 구간**(DB·LLM·**네트워크**)** 여부를 검증함으로써

서비스 출시 전에 전반적인 안정성을 점검하는 것을 목적으로 한다.

III. 테스트 시나리오

시나리오	설명
1) 로그인 **API** 성능 측정	`/api/auth/login` API의 평균·최대 응답속도 측정 (단일 사용자 및 부하 환경)
2) 냉장고 재료 기반 레시피 추천 **API**	`/api/me/recommendations` API 처리 속도 측정 (OpenAI LLM 호출 포함)
3) 커뮤니티 게시글 등록	`/api/events/{event_id}/posts` API 응답 속도 측정. 이벤트당 게시글 3개 제한 조건 반영
4) DB 부하 테스트	- 로그인/사용자 패턴: user_info - 레시피 추천/선택 패턴: recommend_recipe, selected_recipe - 게시글(커뮤니티) 패턴: board - 행동 로그 패턴: board_likes
5) 동시 사용자 부하 테스트	로그인 API를 기준으로 50명, 100명, 200명 부하 테스트

IV. 테스트 결과

IV.1 API 응답속도

기준: 실제 EC2(t3.micro)에서 ApacheBench로 측정한 결과

API	평균 응답속도	최대 응답속도	비고

로그인	**638ms** **(**약 **0.64**초**)**	1837ms (약 1.84초)	정상, ab(`n=1000, c=100`)
레시피 추천	**1779ms** **(**약 **1.78**초**)**	16852ms (약 16.85초)	LLM 호출 포함 시 증가, ab(n=300, c=30)
게시글 등록	**2.45ms** **(**약 **0.002**초**)**	3ms (약 0.003초)	이벤트당 3개 제한 검증을 포함해도 응답 속도는 매우 빠른 수준, ab(n=3, c=1)

요약

- 로그인 API는 평균 약 **0.64**초, 최대 약 **1.8**초로, 인증 절차와 DB 조회를 포함한 응답 속도가 실사용에 무리가 없는 수준으로 확인되었다.
- 레시피 추천 API는 **LLM** 호출이 포함되어 평균 응답속도가 약 **1.78**초로 상대적으로 길지만, 중앙값이 약 **89ms** 수준이라 대부분의 요청은 매우 빠르게 처리되고, 소수의 LLM 응답 지연이 최대 약 16초까지 발생하면서 평균을 끌어올리는 패턴을 보인다.
- 게시글 등록 API는 비즈니스 룰(이벤트당 3개 제한) 검증을 포함하더라도 약 2~3ms 수준으로 매우 빠르게 처리된다.

IV.2 레시피 추천 모델 단위 성능 테스트 **(LLM** 구간 분석**)**

테스트 개요

- 대상 API: /api/me/recommendations
  (냉장고 재료 기반 레시피 추천, OpenAI LLM 호출 포함)
- 방법:
  - 동일 사용자(devtest) 기준
  - 냉장고 재료 상태를 변경하면서 10회 호출
  - 코드 내부에 타이머를 삽입하여
    - LLM 구간(adapt_recipes_json) 처리 시간
    - 전체 추천 처리 시간(recommend_json 전체) 측정

측정결과 (10회 호출 기준)

구분	최소	최대	평균
LLM 구간 시간	7.07초	18.26초	약 14.23초
전체 추천 처리 시간	7.76초	18.82초	약 14.93초

- 전체 응답 시간 중 **대부분(약 95%)**이 **LLM** 구간에서 소요됨
  (LLM 평균 14.23초 / 전체 평균 14.93초 기준).

해석

- 재료가 변경되어 새로운 **LLM** 호출이 발생하는 추천 시나리오에서
  평균 약 **15**초 내외의 응답 시간이 소요되는 것으로 확인되었다.

- 후보 레시피 조회/후처리(DB 작업 등)는 전체의 일부(약 1초 미만)만 차지하며,
  성능 병목은 거의 전적으로 **LLM** 호출 구간에 집중되어 있다.

- 별도 실험에서 단일 호출이 약 **30**초 수준까지 지연되는 **outlier**도 관측되었으며,
  이는 외부 LLM 응답 지연 및 네트워크 상태에 따른 영향으로 해석할 수 있다.

IV.3 동시 사용자 처리 **(**부하 테스트**)**

동시 사용자 수	**ab** 설정 **(n, c)**	평균 응답시간	성공률	서버 상태
50명	**n=500, c=50**	**219ms**	100%	안정적
100명	**n=1000, c=100**	**638ms**	100%	응답 시간은 증가하지만, 여전히 허용 가능한 수준. 에러 없음
200명	**n=2000, c=200**	**10329ms**	90%	응답 지연 급증 및 일부 요청 실패 발생.

해석

- 동시 **50**명·**100**명까지는 실패 없이 안정적으로 요청을 처리하며, 평균 응답시간도 1초 이내로 유지된다.
- 동시 **200**명에서는 평균 응답시간이 약 10초를 넘어가고 Non-2xx 응답(길이 불일치 등) 195건이 발생하여, 현재 인프라와 구조에서의 실질적인 한계로 볼 수 있다.
- 따라서 현재 구성 기준으로는 **"동시 100명 수준까지는 안정적, 200명 이상에서는 확장 전략 필요"**로 요약할 수 있다.

IV.4 서버 자원 사용량

도구: EC2 내부에서 `htop` 실행
상황: 로그인·추천API 부하 테스트 수행 중 관측

지표	사용량
CPU 사용량 평균	2-3%
CPU 피크	3% (레시피 추천 API 부하 시점 기준)
메모리 사용량 평균	**54~55%** (약 540MB / 996MB 사용)
네트워크 대역폭 사용량	수 Mbps 미만 수준으로 추정, 이번 테스트 범위에서는 자원 병목 미발견

해석

- CPU와 메모리는 전반적으로 여유 있는 상태로, EC2 t3.micro 환경에서도 기본적인 트래픽 수준에서는 자원 부족 문제가 발생하지 않았다.
- 레시피 추천 API 테스트 시에도 CPU 사용률은 크게 오르지 않아, 병목이 **EC2 CPU/**메모리가 아닌 외부 **LLM(OpenAI)** 응답 지연 및 네트워크 왕복 시간 쪽에 더 가깝다는 점을 확인했다.

IV.5 DB 성능 테스트

작업	실제 실행 쿼리 (요약)	처리 시간(초)	처리 시간(ms)	평가
`user_info` 단건 조회	`SELECT * FROM user_info WHERE user_id = 'testuser01';`	0.016 s	**16 ms**	매우 빠름
`recommend_recipe` 최근 20개	`SELECT * FROM recommend_recipe WHERE user_id = 'testuser01' ORDER BY created_at DESC LIMIT 20;`	0.015 s	**15 ms**	매우 빠름
`selected_recipe` INSERT	`INSERT INTO selected_recipe (...) VALUES (...);`	0.016 s	**16 ms**	빠름 (쓰기 부담 거의 없음)
`board` INSERT	`INSERT INTO board (...) VALUES (...);`	0.000 s	**≈ 0 ms**	매우 빠름 (즉시 처리)
`event_result` JOIN 조회	`SELECT er.*, e.event_name, u.nickname FROM event_result er JOIN event e ... JOIN user_info u ... WHERE er.event_id = 1 LIMIT 50;`	0.016 s	**16 ms**	JOIN 2회 포함해도 빠른 수준

요약

- 단건 조회/쓰기, 최근 20건 조회, 2중 JOIN 조회까지 모두 **15~16ms** 수준으로 처리되어, 현재 데이터량과 스키마 기준에서는 **DB** 자체가 병목으로 작용하지 않음을 확인했다.
- 게시글 INSERT는 측정상 0초(≈ 0ms)에 가까운 속도로, 쓰기 비용이 매우 낮은 테이블 구조임을 보여준다.

- 향후 실제 데이터가 많이 쌓인 이후에는 `event_result`, `board`, `recommend_recipe` 중심으로 `EXPLAIN` 및 인덱스 튜닝을 통해 현재 수준의 응답성을 유지하는 것이 중요하다.

V. 성능 개선 방안

개선 항목	내용
LLM 호출 최소화 **(**후보 풀 캐싱**)**	같은 재료·프로필에 대해 LLM이 만든 **후보 레시피 풀(10~15개)**을 일정 시간 캐싱하고, 매 요청마다 그 안에서 최근 노출 레시피를 제외한 **3**개만 섞어서 추천해 LLM 호출 횟수를 줄인다.
API 레벨 캐싱	Redis 등 In-memory 캐시와 프론트 캐시를 활용해 동일 요청에 대한 중복 추천 호출을 최소화한다.
비동기 처리·워커 분리	LLM 호출을 비동기 작업 또는 별도 워커로 분리해, 메인 API는 빠르게 응답하고 추천 결과는 후속으로 전달하는 구조를 검토한다.
LLM 호출 최소화	동시 200명 이상 구간을 대비해 오토 스케일링·로드 밸런서·토큰 발급 로직 최적화로 피크 트래픽 대응력을 높인다
DB 인덱스·쿼리 점검	`board`, `board_likes`, `event_result` 중심으로 **EXPLAIN**·복합 인덱스를 적용해 조회·집계 성능을 개선한다.
모니터링 고도화	`CloudWatch·APM` 등을 활용해 **`LLM`** 응답시간·**`DB`** 쿼리·에러율을 모니터링하고, 병목 구간을 지속적으로 추적한다.

VI. 결론

- 동시 **50~100**명 수준에서는 로그인·추천·게시글 등록 모두 에러 없이 안정적인 성능을 보였다.

- 로그인 API는 동시 **200**명부터 평균 약 **10.3**초와 일부 실패가 발생해, 현재 t3.micro 환경의 실질적인 한계 구간이 확인되었다.

- 레시피 추천 API는 평균 **1779ms(1.78초)**이지만, 중앙값 **89ms**로 대부분은 빠르게 응답하고, **소수의 LLM 지연(outlier)**이 평균을 끌어올리는 구조다.

- CPU 2~3%, 메모리 54~55% 수준으로 서버 자원은 여유가 있으며, **LLM** 호출 최적화·캐싱·스케일아웃·**DB** 튜닝을 적용하면 더 많은 동시 사용자와 빠른 응답속도를 지원할 수 있을 전망이다.

VII. 부록

VII. 부록 **(**구성 예시**)**

- **A. ApacheBench(ab)** 명령어 정리
    - 로그인 API 단일·부하 테스트용 명령어

```
[ec2-user@ip-172-31-41-205 ~]$ ab -n 500 -c 50 -p login.json -T application/json http://43.203.1.85/api/auth/login
This is ApacheBench, Version 2.3 <$Revision: 1923142 $>
Copyright 1996 Adam Twiss, Zeus Technology Ltd, http://www.zeustech.net/
Licensed to The Apache Software Foundation, http://www.apache.org/

Benchmarking 43.203.1.85 (be patient)
Completed 100 requests
Completed 200 requests
Completed 300 requests
Completed 400 requests
Completed 500 requests
Finished 500 requests


Server Software:        nginx/1.27.5
Server Hostname:        43.203.1.85
Server Port:            80

Document Path:          /api/auth/login
Document Length:        218 bytes

Concurrency Level:      50
Time taken for tests:   2.915 seconds
Complete requests:      500
Failed requests:        0
Total transferred:      443500 bytes
Total body sent:        95000
HTML transferred:       109000 bytes
Requests per second:    171.54 [#/sec] (mean)
Time per request:       291.482 [ms] (mean)
Time per request:       5.830 [ms] (mean, across all concurrent requests)
Transfer rate:          148.59 [Kbytes/sec] received
                        31.83 kb/s sent
                        180.42 kb/s total

Connection Times (ms)
              min  mean[+/-sd] median   max
Connect:        0    0   0.3      0       2
Processing:    70  263 138.1    238    1453
Waiting:       69  263 138.1    237    1453
Total:         71  263 138.0    238    1453

Percentage of the requests served within a certain time (ms)
  50%    238
  66%    262
  75%    275
  80%    280
  90%    308
  95%    321
  98%    503
  99%   1303
 100%   1453 (longest request)
```

```
[ec2-user@ip-172-31-41-205 ~]$ ab -n 1000 -c 100   -p login.json   -T application/json   http://43.203.1.85/api/auth/login
This is ApacheBench, Version 2.3 <$Revision: 1923142 $>
Copyright 1996 Adam Twiss, Zeus Technology Ltd, http://www.zeustech.net/
Licensed to The Apache Software Foundation, http://www.apache.org/

Benchmarking 43.203.1.85 (be patient)
Completed 100 requests
Completed 200 requests
Completed 300 requests
Completed 400 requests
Completed 500 requests
Completed 600 requests
Completed 700 requests
Completed 800 requests
Completed 900 requests
Completed 1000 requests
Finished 1000 requests


Server Software:        nginx/1.27.5
Server Hostname:        43.203.1.85
Server Port:            80

Document Path:          /api/auth/login
Document Length:        218 bytes

Concurrency Level:      100
Time taken for tests:   6.426 seconds
Complete requests:      1000
Failed requests:        0
Total transferred:      887000 bytes
Total body sent:        190000
HTML transferred:       218000 bytes
Requests per second:    155.63 [#/sec] (mean)
Time per request:       642.559 [ms] (mean)
Time per request:       6.426 [ms] (mean, across all concurrent requests)
Transfer rate:          134.81 [Kbytes/sec] received
                        28.88 kb/s sent
                        163.68 kb/s total

Connection Times (ms)
              min  mean[+/-sd] median   max
Connect:        0    0   0.7      0      11
Processing:    75  528 147.8    530    1837
Waiting:       72  528 147.9    530    1837
Total:         75  529 147.7    531    1837

Percentage of the requests served within a certain time (ms)
  50%    531
  66%    546
  75%    554
  80%    564
  90%    580
  95%    610
  98%    788
  99%   1493
 100%   1837 (longest request)
[ec2-user@ip-172-31-41-205 ~]$
```

```
[ec2-user@ip-172-31-41-205 ~]$ ab -n 2000 -c 200 -p login.json -T application/json http://43.203.1.85/api/auth/login
This is ApacheBench, Version 2.3 <$Revision: 1923142 $>
Copyright 1996 Adam Twiss, Zeus Technology Ltd, http://www.zeustech.net/
Licensed to The Apache Software Foundation, http://www.apache.org/

Benchmarking 43.203.1.85 (be patient)
Completed 200 requests
Completed 400 requests
Completed 600 requests
Completed 800 requests
Completed 1000 requests
Completed 1200 requests
Completed 1400 requests
Completed 1600 requests
Completed 1800 requests
Completed 2000 requests
Finished 2000 requests


Server Software:        nginx/1.27.5
Server Hostname:        43.203.1.85
Server Port:            80

Document Path:          /api/auth/login
Document Length:        218 bytes

Concurrency Level:      200
Time taken for tests:   103.286 seconds
Complete requests:      2000
Failed requests:        195
   (Connect: 0, Receive: 0, Length: 195, Exceptions: 0)
Non-2xx responses:      195
Total transferred:      1642417 bytes
Total body sent:        380000
HTML transferred:       400797 bytes
Requests per second:    19.36 [#/sec] (mean)
Time per request:       10328.610 [ms] (mean)
Time per request:       51.643 [ms] (mean, across all concurrent requests)
Transfer rate:          15.53 [Kbytes/sec] received
                        3.59 kb/s sent
                        19.12 kb/s total

Connection Times (ms)
              min  mean[+/-sd] median   max
Connect:        0  650 4083.4      0   32338
Processing:    75 4470 8474.2   1091   60006
Waiting:       70 4470 8474.2   1091   60006
Total:         75 5120 9698.9   1097   92207

Percentage of the requests served within a certain time (ms)
  50%   1097
  66%   1170
  75%   3418
  80%  10732
  90%  13161
  95%  20679
  98%  35056
  99%  60001
 100%  92207 (longest request)
```

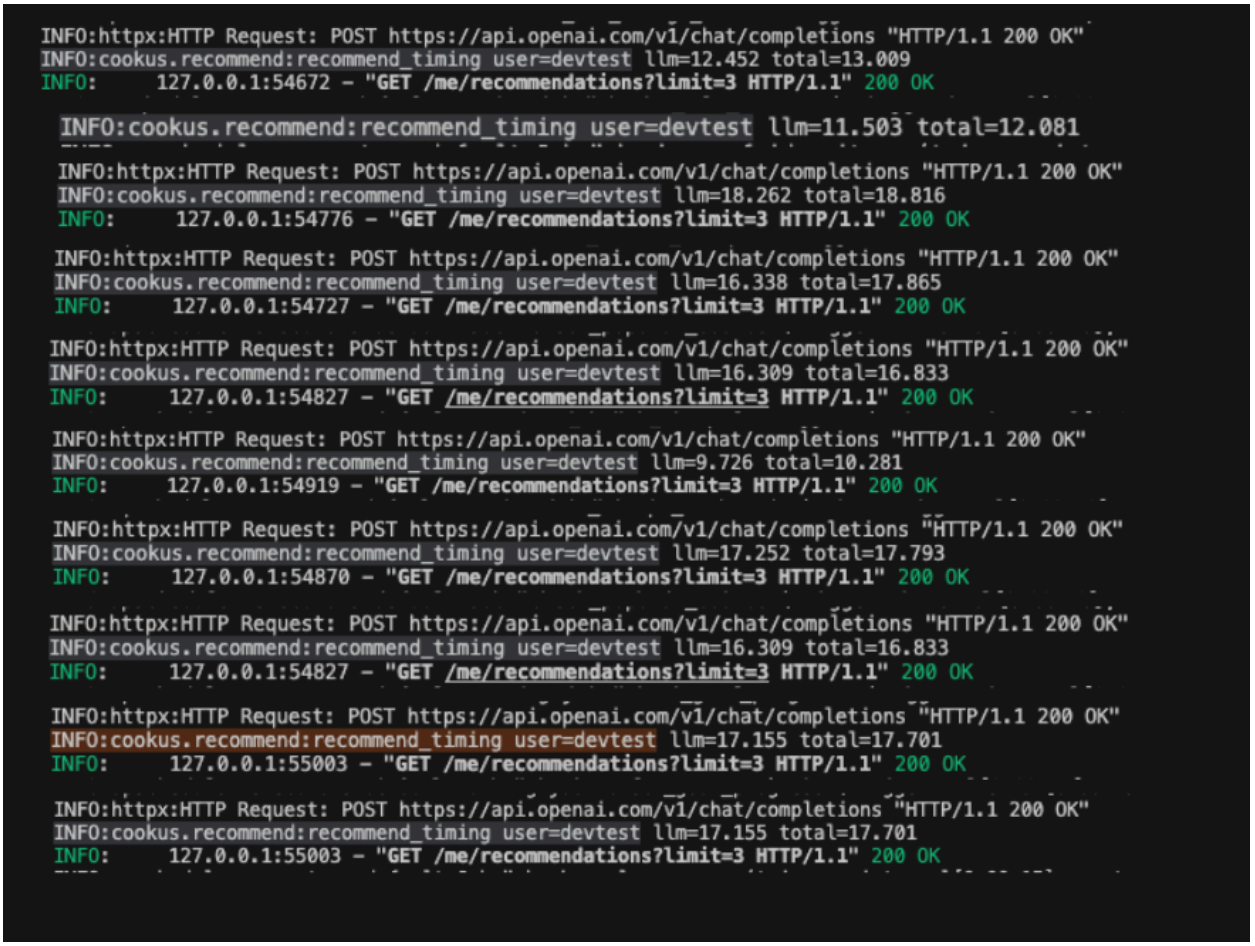
- 레시피 추천 API 부하 테스트(`n=300, c=30`)

```
# 3) ab 실행
ab -n 300 -c 30 \
  -H "Authorization: Bearer $TOKEN" \
  http://43.203.1.85/api/me/recommendations
eyJhbGciOiJIUzI1NiIsInR5cCI6IkpXVCJ9.eyJzdWIiOiJ6eGN2Ym4iLCJpYXQiOjE3NjM0MjQ0MTQsImV4cCI6MTc2MzQyNjIxNH0.wlBLoJE_1JtZBd8CiwyiwfmTh81L6TrilkB0fygVzaI
This is ApacheBench, Version 2.3 <$Revision: 1923142 $>
Copyright 1996 Adam Twiss, Zeus Technology Ltd, http://www.zeustech.net/
Licensed to The Apache Software Foundation, http://www.apache.org/

Benchmarking 43.203.1.85 (be patient)
Completed 100 requests
Completed 200 requests
Completed 300 requests
Finished 300 requests


Server Software:        nginx/1.27.5
Server Hostname:        43.203.1.85
Server Port:            80

Document Path:          /api/me/recommendations
Document Length:        2118 bytes

Concurrency Level:      30
Time taken for tests:   17.789 seconds
Complete requests:      300
Failed requests:        0
Total transferred:      680700 bytes
HTML transferred:       635400 bytes
Requests per second:    16.86 [#/sec] (mean)
Time per request:       1778.949 [ms] (mean)
Time per request:       59.298 [ms] (mean, across all concurrent requests)
Transfer rate:          37.37 [Kbytes/sec] received

Connection Times (ms)
              min  mean[+/-sd] median   max
Connect:        0    0   0.1      0       1
Processing:    61  146 967.8     89   16852
Waiting:       61  146 967.8     89   16851
Total:         61  147 967.8     89   16852

Percentage of the requests served within a certain time (ms)
  50%     89
  66%     95
  75%     97
  80%     99
  90%    107
  95%    115
  98%    154
  99%    160
 100%  16852 (longest request)
```

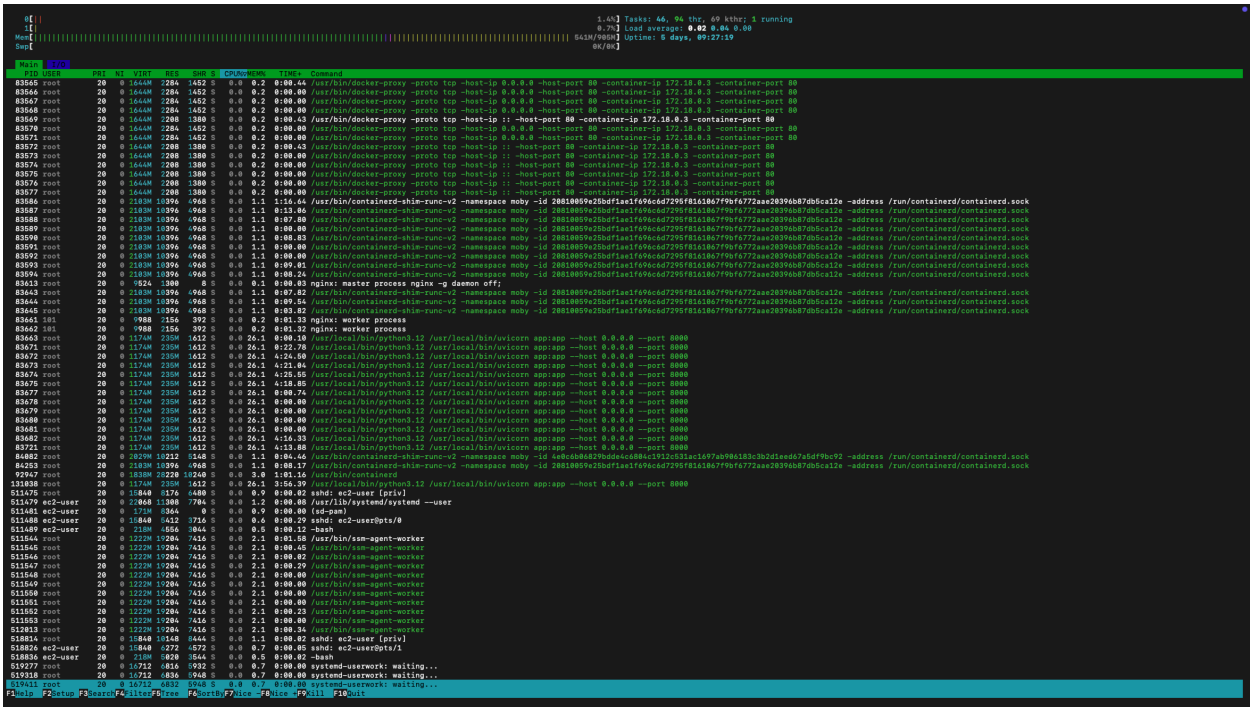
- 게시글 등록 API 테스트(n=3, c=1)

```
[ec2-user@ip-172-31-41-205 ~]$ ab -n 3 -c 1 \
  -p post.json \
  -T application/json \
  -H "Authorization: Bearer $TOKEN" \
  http://43.203.1.85/api/events/3/posts
This is ApacheBench, Version 2.3 <$Revision: 1923142 $>
Copyright 1996 Adam Twiss, Zeus Technology Ltd, http://www.zeustech.net/
Licensed to The Apache Software Foundation, http://www.apache.org/

Benchmarking 43.203.1.85 (be patient).....done


Server Software:        nginx/1.27.5
Server Hostname:        43.203.1.85
Server Port:            80

Document Path:          /api/events/3/posts
Document Length:        26 bytes

Concurrency Level:      1
Time taken for tests:   0.007 seconds
Complete requests:      3
Failed requests:        0
Non-2xx responses:      3
Total transferred:      555 bytes
Total body sent:        1305
HTML transferred:       78 bytes
Requests per second:    408.61 [#/sec] (mean)
Time per request:       2.447 [ms] (mean)
Time per request:       2.447 [ms] (mean, across all concurrent requests)
Transfer rate:          73.82 [Kbytes/sec] received
                        173.58 kb/s sent
                        247.40 kb/s total

Connection Times (ms)
              min  mean[+/-sd] median   max
Connect:        0    0   0.1      0       0
Processing:     2    2   0.6      2       3
Waiting:        2    2   0.6      2       3
Total:          2    2   0.7      3       3
WARNING: The median and mean for the total time are not within a normal deviation
        These results are probably not that reliable.

Percentage of the requests served within a certain time (ms)
  50%      2
  66%      2
  75%      3
  80%      3
  90%      3
  95%      3
  98%      3
  99%      3
 100%      3 (longest request)
```

- **b. LLM** 구간분석 스크린샷



- **C. htop** 스크린샷
    - 로그인 부하 시 CPU/MEM 상태



- 레시피 추천 부하 시 CPU/MEM 상태