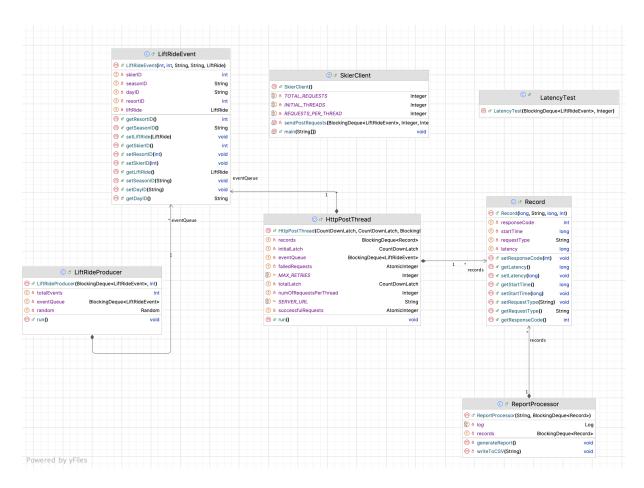
git repo URL

• https://github.com/Syl-Ying/Skier

Client Design



• A Single Producer

- LiftRideEvent class: a lift ride event.
- LiftRideProducer class: generate 200k random LiftRideEvent instances and store them in a thread-safe queue for the worker threads to consume.

A1 1

• Multithreading Consumers:

- HttpPostThread class: a Runnable class to send given numbers of requests. Each POST request will retry up to 5 times upon receiving 4XX or 5XX errors.
- SkierClient class: the main class
 - First use a single thread to generates 200k events.
 - Then 32 initial threads, each sending 1000 POST requests, will start.
 - Once those threads finish, 168 threads will be created until all 200K requests are sent.
- **Performance Metrics**: Track the number of successful and unsuccessful requests, the total runtime, and calculate throughput (requests per second).
 - Record class: a request's status code, starting time...
 - ReportProcessor class
 - First calculate related metrics and print to terminal.
 - Then output a CSV file with all the records

Little's Law Throughput Predictions

• Plot of throughput over time (5 points)

Testing:

1 request is sent with 1 thread, total time is 3 ms.

10k requests are sent using 1 thread, total time is 5 ms.

W - Average response time / Latency: 3ms

L - Number of requests sent: 10,000

 $\lambda = L / W = 10,000 / (3 / 1000) = 3,333,000$ rps. so expected throughput is 3,333,000.

Actual:

200k requests are sent using 200 threads.

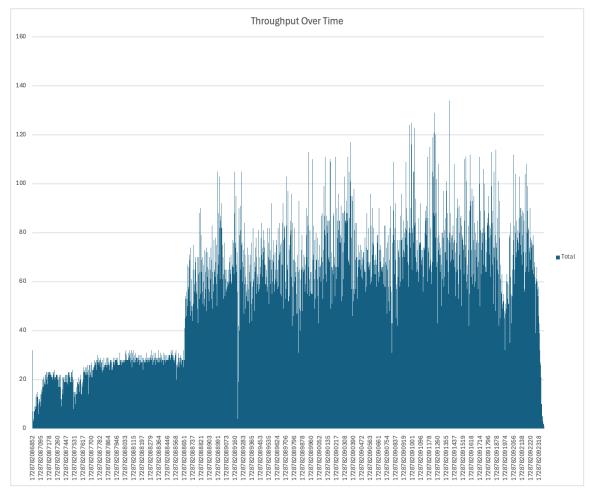
W: Mean response time 5 ms.

N: I have 200 client threads and 200 default Tomcat threads, and hence the maximum is 200 concurrent requests.

Prediction $\lambda = 200 / 0.005 = 40,000 \text{ rps}$

Actual λ = Actual Requests / Actual Time = 200k / 8.515 = 23,488 rps

The difference is caused by waiting for one of the 32 threads to finish before starting the remaining 168 threads in actual settings. The prediction assumes 200 threads are started simultaneously.



Y axis: requests numbers at the specific time. X axis: time

A1 3

Client (Part 1)

a screenshot of output window with your wall time and throughput. include the client configuration in terms of the number of threads used (print it out in your output window).

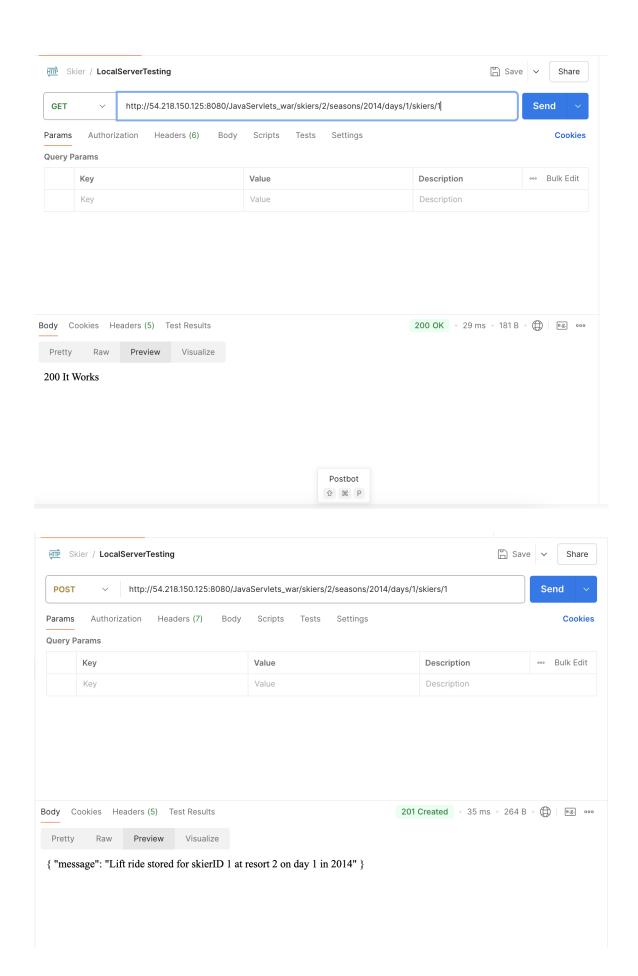
Client (Part 2)

• run the client as per Part 1, showing the output window for each run with the specified performance statistics listed at the end.

EC₂

• screenshots to show you actually send your requests to the server on EC2 instance: Postman testing page showing the Url, and/or the EC2 instance page.

A1 4



5