Informatics Institute of Technology

In Collaboration With

University of Westminster, UK

University of Westminster, Coat of Arms

Analysing and Modelling the Latency and Accuracy Trade-offs in Rate Limiting on API-Gateway

A dissertation by

Mr. Krishnamoorthy Caucidheesan

W1790009 | 20191126

Supervised by

Dr. Malith Jayasinghe

May 2023

Submitted in partial fulfilment of the requirements for the

BSc (Hons) Computer Science degree at the University of Westminster.

ABSTRACT

The rate limiting service in API gateways controls request entry by throttling the requests with a boundary. The rate limiting accuracy determines how efficiently rate limiting service works and on whether it allows requests within the throttle count or whether it allows more than the throttle count. Latency on the other hand, is the round-trip time of a particular API call. The accuracy of rate limiting service is defined using spillover error percentage. It is the percentage of requests that rate limiting service allows more than the throttle count.

Requests must be sent to the rate limiting service in order to rate limit requests in API gateway. The rate limiting service decides whether the requests arrived must be throttled or not. The time taken for the requests to be decided by the rate limiting service adds an additional latency to the round-trip time of the requests. This additional latency can be controlled by introducing a time-out. However, this could result in a degradation in the accuracy of rate limiting. The objective of the project has two folds. First, the author investigates the relationship between accuracy and latency under different workload conditions and timeouts. Secondly, develop a set of machine-learning models that can estimate the accuracy & latency given the relevant parameters.

The analysis of accuracy and latency tradeoff is completed by deriving conclusions based on the observation of results from performance tests conducted. The conclusions state there is a positive relationship between accuracy and latency and the tradeoff was identified. Voting regressor ensembling technique combining the prediction of multivariate linear regressor model and random forest regressor model was used to predict the accuracy and latency trade off. The models have R2 score of 0.85 ± 0.04 with low Mean Absolute Error (< 8.00) and low Mean Squared Error (< 14.00).

Keywords: Spillover error percentage, Latency, Trade-off, Rate limiting service, API gateways, Multivariate Linear Regression, Random Forest Regression

Subject Descriptors:

- Theory of Computation → Models of computation → Concurrency → Parallel computing models.
- Computing methodologies → Machine learning → Learning paradigms → Supervised learning → Supervised learning by regression.