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

Krishnamoorthy Caucidheesan[1*], Guhanathan Poravi[2]

[1]*Computer Science and Engineering, University of Westminster, London, UK, w1790009@westminster.ac.uk*

[2]*Dept. of Computing, Informatics Institute of Technology, Colombo, Sri Lanka, guhanathan.p@iit.ac.lk*

The rate limiting service in API gateways controls request entry by throttling requests with a boundary. The rate limiting accuracy determines how efficiently it works and whether it allows requests within the throttle count or exceeds 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 error calculation for the 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 the API gateway. The rate limiting service decides whether the incoming requests must be throttled. The time taken for the requests to be decided by the rate limiting service adds additional latency to the round trip time of requests. This additional latency can be controlled by introducing a timeout. However, this could result in a degradation in the accuracy of the rate limiting. This paper investigates the particular problem and models the relationship between accuracy and round-trip latency. The findings of this research address the analysis of the accuracy and latency trade-off with respect to the parameters influencing them, and also address the prediction outcome using random forest regressor and present key findings.

Keywords: *accuracy, API gateways, latency, random forest regressor, rate limiting service, trade-off*