# Scaling and Performance of the system

How I would deal with scaling and performance would differ on case by case, for example should the API but in production and suddenly get hit by a sudden long increase of valid traffic best would likely be scale the application out, whether it be using autoscaling that is offered by many cloud platforms / Kubernetes or if auto scaling is not an option you can always look at commissioning more servers but that’s not always an option.

If the scaling does not help then it would be time to get the hands dirty and look at the code to see if there is any optimizations that can be done to improve performance whether it be in the code that is execute or the database queries that are executed.

Ideally before API’s hit production they should go through a series of load tests and profiling to see what its limitations is which could save a lot of pain, something that we currently do at the company I work is we implemented AOP to do logging for all incoming traffic and outgoing responses, those metrics are then exported to elastic and finally shipped to Grafana where we build our monitoring and alerting where we can pickup issues quite quickly.