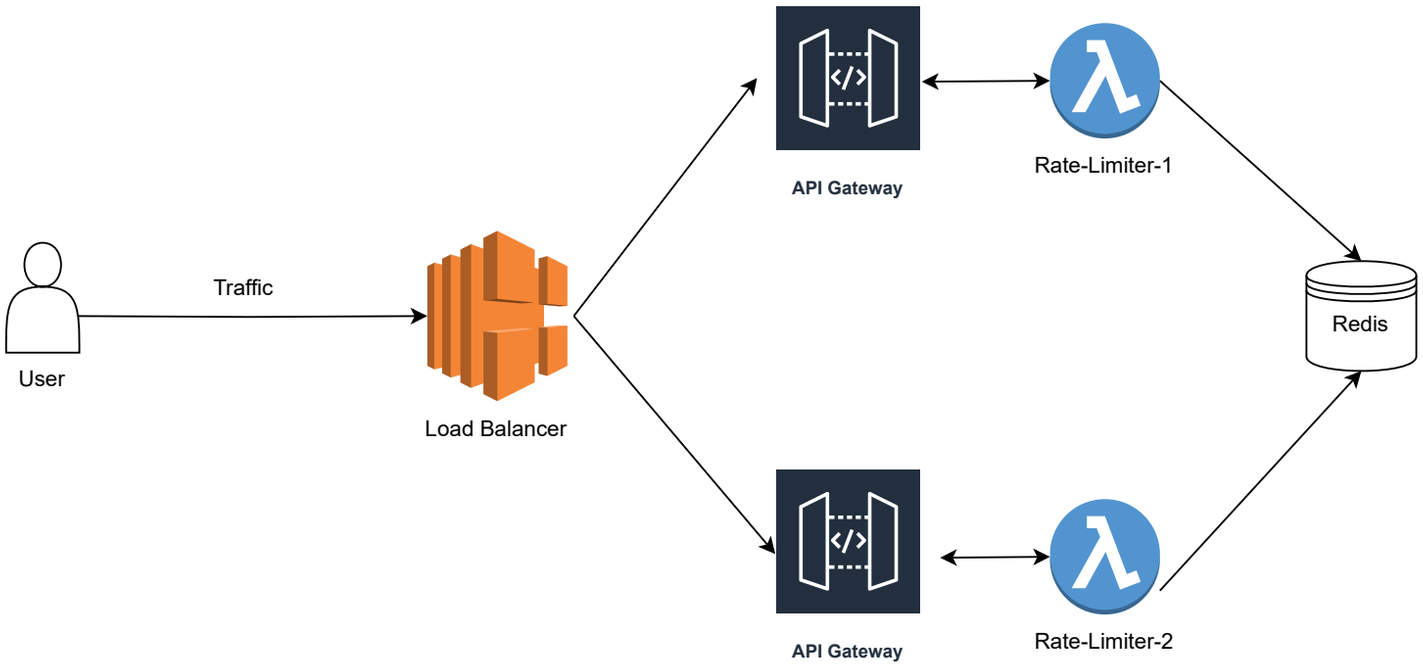

This is API Gateway design for high availability.

- Have one Load Balancer in front of two API Gateway instances. It coordinates traffic from user to gateways using round robin scheduling.
- Implement Rate-limiter per API Gateway. The Rate-limiter operates independently. So, each Gateway must have a different Rate-limiter's name (can specify by `LIMITER_NAME` environment variable or argument when call class method).
- Implement Redis to storage request for Rate-limiter.