

Ограничение скорости и регулирование — продвинутые темы в проектировании REST API, которые могут помочь управлять трафиком API и защитить его от чрезмерного использования или злоупотребления. Если слишком много запросов будет идти к API, сервер может просто не справиться.

Рассмотрим основные практики:

1. Регулирование на уровне количества вызовов API

Можно настроить сервер так, чтоб он ограничил количество запросов, которые могут быть сделаны к API в течение какого-то времени. Ограничение количества запросов может быть основано на различных критериях, таких как количество запросов в целом в час, или количество запросов от определенного IP-адреса или конкретного пользователя (допустим, вы узнаете его по x-api-key). Ограничение количества запросов позволяет предотвратить перегрузку API слишком большим трафиком или защитить его от использования в злонамеренных целях, например, от DDOS атак.

Как это реализуется?

Когда клиент превышает лимит запросов, вы отвечаете ошибкой 429 Too Many Requests и заголовками:

- X-Rate-Limit-Limit — количество разрешенных запросов в текущем периоде (всего)
- X-Rate-Limit-Remaining — количество оставшихся запросов в текущем периоде
- X-Rate-Limit-Reset — количество секунд, оставшихся в текущем периоде

Так клиент сможет понять, через какое время он может возобновить запросы, и как в целом ему выстроить работу с вами на основе количества разрешенных запросов в принципе.

2. Ограничение скорости обработки запросов на уровне вашего приложения, когда запрос уже получен по API

Часто используется для защиты внутренних систем от перегрузки при внезапном наплыве запросов или для приоритетной обработки срочных запросов. Может быть реализовано различными способами, например, путем добавления задержек между запросами или пакетной обработки запросов.

Как это реализуется?

Разработчики приложения используют конкретные алгоритмы, которые нужно рассматривать для каждого случая отдельно.