***Rate Limiting Approaches***

**1) Fixed Window Counter**

* Splits time into fixed periods (e.g., 1-second intervals).
* Counts how many requests a user makes in each period.
* Allows up to a set number of requests (e.g., 5 per second).
* Resets the count at the start of each new period.

**2) Token Bucket**

* Each user has a bucket that holds a certain number of tokens.
* Tokens are added at a steady rate (e.g., 5 per second).
* Every request removes one token from the bucket.
* If no tokens are left, requests must wait until new tokens are added.
* Helps spread requests evenly and prevents sudden traffic spikes.

***Trade-Offs Between Approaches***

**1) Fixed Window Counter**

* Simple and easy to set up.
* Uses little memory since it only tracks request counts.
* Can cause short bursts of extra traffic when limits reset (e.g., making 5 requests at the end of one window and 5 more at the start of the next).
* Not flexible for changing traffic patterns.

**2) Token Bucket**

* Helps control traffic smoothly by refilling tokens gradually.
* Allows users to save unused tokens for later use.
* Harder to implement because it requires constant tracking of token levels.
* Needs extra processing power to manage token updates.