Rate Limiting Algorithms

Rate limiting is a technique to control the rate of requests sent or received by a system. This document covers common algorithms:

1. Fixed Window
   - Simple counter reset at fixed intervals
   - Pros: Easy to implement
   - Cons: Burst at window boundaries

2. Sliding Window
   - Track individual request timestamps
   - Pros: Smooth rate limiting
   - Cons: Higher memory usage

3. Token Bucket
   - Tokens added at fixed rate
   - Each request consumes a token
   - Pros: Allows controlled bursts

4. Leaky Bucket
   - Requests processed at fixed rate
   - Queue for overflow
   - Pros: Smooth output rate