Rate Limiting

Inhaltsverzeichnis

Rate Limiting begrenzt, wie viele Anfragen ein Client in einem Zeitraum stellen darf. Es schützt APIs und Login-Endpunkte vor Missbrauch, Überlast und Brute-Force.

1 Funktionsweise

Zähler oder Token-Buckets erfassen Requests pro IP, User oder API-Key. Bei Überschreitung folgen HTTP 429, Delays oder Blocks.

- **Fenster:** z. B. Requests/Minute
- **Scope:** IP, Identität, Endpunkt
- **Antwort:** 429 inkl. Retry-Hints

2 Typische Einsatzbereiche

- Public APIs und Login/MFA-Endpunkte
- Schutz vor Scraping und Credential Stuffing
- Fairness in Multi-Tenant-Systemen

3 Wichtige Hinweise

- Zu strenge Limits treffen legitime Spitzen.
- Verteilte Limiter brauchen gemeinsamen State (Redis o. Ä.).
- Bypass für interne Health-Checks klar regeln.

4 Fazit

Rate Limiting ist einfacher, wirksamer Schutz für APIs und Auth-Flows – fein justiert nach Nutzungsmuster statt pauschal.