

Kritik Endpoint'ler İçin Kaynak/Rate Limit & Throttling Planlama Şablonu — Yazılım / Sunucu ve Güvenlik (v1.0)

Asset Amaç: Bu şablon, sistem performansını ve altyapı güvenliğini eşzamanlı olarak koruyan katmanlı bir limitleme modeli kurmak için hazırlanmıştır. İşletim sistemi (OS), web sunucu (Nginx/HAProxy), uygulama çalışma zamanı ve API Gateway katmanlarındaki kaynak limitleri ile kritik endpoint'lerin rate limit / throttling profillerini tek bir strateji dökümanında birleştirir. Amaç; ani trafik artışlarında, Kaba Kuvvet (Brute-Force) ve DoS/DDoS saldırılarında tam kesinti (*outage*) yaşamak yerine sistemin kontrollü yavaşlamasını (*graceful degradation*) ve istemcilere standart yanıtlar sunarak ayakta kalmasını sağlamaktır.

Kim Kullanır?: Backend Geliştiriciler, DevOps / SRE Ekipleri, Platform Mühendisliği Birimleri ve Otel / B2B Ürün Ekipleri.

Nasıl Kullanılır?

- Servislerinizdeki tüm kritik endpoint'lerin envanterini çıkararak risk ve suiistimal (*abuse/overload*) seviyelerini belirleyin.
- İşletim sisteminden API katmanına kadar eşik değerlerini belirleyin; isteklerin reddedilme veya kuyruğa alınma (*throttling & queue*) davranışlarını tanımlayın.
- Observability panellerine alarm eşiklerini ekleyin ve gerçek trafik loglarına (*access logs*) bakarak limitleri periyodik olarak optimize edin.

TEMPLATE — Kritik Endpoint'ler İçin Kaynak/Rate Limit & Throttling Planı

1) Endpoint Envanteri ve Risk Matrisi

Endpoint	Tip	Kritik mi?	Risk (Abuse / Overload)	Sahip Ekip
/login	Auth	Evet (Kritik)	Yüksek (Brute-force / Credential Stuffing)	Security / Auth Ekibi
/search	Query	Evet	Orta-Yüksek (DB CPU spike / Unindexed search)	Search / Core API

/booking	Transaction	Evet (Kritik)	Yüksek (Race condition / Inventory lock)	Booking / PMS Ekibi
/report	Heavy Query	Hayır	Yüksek (Memory exhaustion / Long lock)	Reporting / Analytics
/export	Batch	Hayır	Kritik (Disk I/O / OOM Kill)	Platform / Infrastructure

2) Katman Bazlı Limit Planı

Katman	Limit Tipi	Standart Değer / Eşik	Aksiyon / Yönlendirme
OS (İşletim Sistemi)	CPU / RAM / Connection	max_open_files: 65535, sysctl net.core.somaxconn: 4096	TCP paket birikmesini önleme, FD tükenmesini engelleme
Web Server (Nginx)	Conn / Timeouts	limit_conn_zone: 50 conn/ip, keepalive_timeout: 15s	Mikro burst'leri göğüsleme, boş bağlantıları düşürme
Uygulama Runtime	Worker / Thread / Pool	DB Connection Pool: Max 50, Puma/Gunicorn Worker: 8	Veritabanı bağlantı havuzunun (pool) kilitletmesini engelleme
API Gateway	IP / User / Token	rate_limit: 100 req/min (Global IP), burst: 20	HTTP 429 Too Many Requests yanıtı ve Retry-After başlığı

3) Endpoint Profilleri ve Throttling Davranışı

Endpoint	Rate Limit Bazları	Throttling / Kuyruk Tipi	Graceful Davranış (Degradation)	Notlar
/login	IP + Kullanıcı Adı	Drop / Block (Kuyruklama yok)	HTTP 429 + Retry-After: 60s	CAPTCHA tetikleme eşiği: 5 hatalı deneme
/search	Token / Session ID	Leaky Bucket (Kısmi kuyruk)	Önceden ön belleğe alınmış (cached) popüler veriyi dönme	ElasticSearch önünde Redis cache zorunlu
/booking	User ID + Tenant ID	Priority Queue (Sıkı önceliklendirme)	HTTP 503 + Standart "Yoğunluk Nedeniyle Bekleyiniz" ekranı	Odaları geçici kilitleme süresi max 5 dk
/report	Tenant ID	Rate Throttling (Hız yavaşlatma)	Ağır sorguları okuma replikasına (Read-Replica) yönlendirme	Aynı anda max 2 canlı rapor sorgusu
/export	Token + Tenant ID	Async Task Queue (RabbitMQ/Celery)	HTTP 202 Accepted + Job ID verip arka planda işleme	Yanıt: "İşlem başlatıldı, indirme bağlantısı e-posta ile iletilecektir."

4) İzleme, Alarm ve Sürdürülebilirlik

- Dashboard Metrikleri:

- **p95 / p99 Latency:** Endpoint bazlı yanıt sürelerinde $t > 1500\text{ ms}$ aşılması.
- **HTTP Status Oranları:** 429 Too Many Requests ve 5xx Server Error artış oranları.
- **Queue Depth:** Arka plan iş kuyruklarının (Celery/RabbitMQ) birikme miktarı (> 500 job).
- **Alarm Eşikleri ve Runbook:**
 - *Eşik:* 1 dakika içerisinde bir IP'den /login adresine 50'den fazla 429 dönmesi $\rightarrow$ Otomatik IP Ban (WAF Rule).
 - *Eşik:* API Gateway genelinde 503 oranı $\geq 5\%$ 'i aşarsa $\rightarrow$ Auto-Scaling Tetikleme + Non-critical Endpoint Disable.
- **180 Günlük Gözden Geçirme Takvimi:**
 - Her 6 ayda bir trafik logları analiz edilerek sezonluk yoğunluklara (Örn: Otel erken rezervasyon dönemi) göre rate-limit threshold'ları güncellenir.

5) Denetim ve Kontrol Listesi (Checklist)

- **[] Endpoint Sınıflandırması Yapıldı:** Tüm API rotaları kritiklik ve kaynak tüketim türüne göre etiketlendi.
- **[] Export / Heavy İşler Asenkronlaştırıldı:** Uzun süren veri indirme ve raporlama işleri HTTP senkron akıştan çıkarılıp mesaj kuyruklarına alındı.
- **[] Booking / Login Önceliklendirildi:** Gelir getiren veya güvenlik riski taşıyan rotalara özel kaynak izolasyonu uygulandı.
- **[] 429 / 503 Yanıtları ve Retry Stratejisi Netleştirildi:** İstemcilere standart JSON hata formatı ve Retry-After header bilgisi dönülüyor.
- **[] İzleme ve Alarm Seti Kuruldu:** Grafana/Prometheus üzerinde rate-limit ihlalleri ve CPU/RAM doyumluk göstergeleri canlıya alındı.

Deliverables (Teslim Edilecek Çıktılar)

- **Kritik Endpoint Limit & Throttling Mimarisi Planı (v1.0):** Katmanlı limitleme matrisini ve endpoint davranışlarını belirten doküman.
- **API Gateway & Web Server Limit Konfigürasyon Seti:** Nginx / Kong / AWS WAF için hazır rate-limit kural kodları.
- **Graceful Degradation & Queue Yönetim Runbook'u:** Trafik patlaması anında devreye girecek acil durum ve asenkron işleme rehberi.

RATE LIMIT & THROTTLING CHECKLIST

- ✓ Kritik Endpoint'ler Sınıflandırıldı (Login/Booking)
- ✓ Ağır İşler (Export) Kuyruğa Alınıyor
- ✓ IP/Kullanıcı Bazlı Rate Limit Tanımlı
- ✓ Throttling ile Kontrollü Yavaşlatma Aktif
- ✓ Limit Aşımında Net Hata Mesajı (429+Retry)
- ✓ Graceful Degradation Kurgulandı
- ✓ 429 ve Queue Depth İzleniyor

[Checklist / Framework Card]

Backend geliştiricileri, SRE uzmanları ve platform mimarları için tasarlanmış, 1200x1200px (1:1) kare formatında yüksek çözünürlüklü bir Performans ve Rate Limit Mimari Kartı. Koyu karbon gri ve amber/siber turuncu tonlarındaki teknolojik zemin üzerinde; API Gateway trafik vanası, Token Bucket algoritması görseli, HTTP 429/503 yanıt göstergesi ve kuyruk (queue) akış şemaları yer alıyor. Kartın üzerinde; "Katmanlı OS / Web / App / API Limitleri", "Kritik Endpoint Throttling & Priority Queues", "Graceful Degradation & Asenkron Export" ve "p95/p99 & 429 Alarm Dashboard" teslimat kalemi başlıkları, yanlarında parlak onay işaretleriyle scannable ve kurumsal bir framework-card formatında sergileniyor.