

# Rate Limit Spike Response

INTERNAL — Helix Engineering / CS Operations

## WHEN THIS RUNBOOK APPLIES

A customer or set of customers is generating API traffic that exceeds platform-wide capacity.  
Symptoms: API p99 latency above 2s, increased 503s, or the rate limiter shedding traffic at the LB.

### STEP 1 — Identify the source

Check the API analytics dashboard, sorted by workspace requests-per-minute. Identify any workspace exceeding 5x its plan's documented limit. Common cause: a customer's bot or integration in a tight loop.

### STEP 2 — Reach out to the customer

Email the workspace owner with the request count, peak RPM, and the API keys involved. Most customers respond within 15 minutes when alerted.

### STEP 3 — Apply temporary throttle

If the spike is causing platform-wide impact, apply a temporary per-workspace cap at 50% of the offending rate. This is a tighter limit than their plan allows; document this in the audit log and notify the customer.

### STEP 4 — Block specific keys

If a single API key is responsible (a runaway script), rotate-and-revoke that key from the admin panel. The customer's other keys continue to work.

### STEP 5 — Capacity review

If the spike is from legitimate traffic and the platform cannot handle current load, escalate to capacity planning. Do not revoke keys for customers within their documented plan limits.

## POST-INCIDENT

If a customer was throttled below their documented plan limit, follow up within 24 hours and offer a credit. Document the incident in the customer's account record.