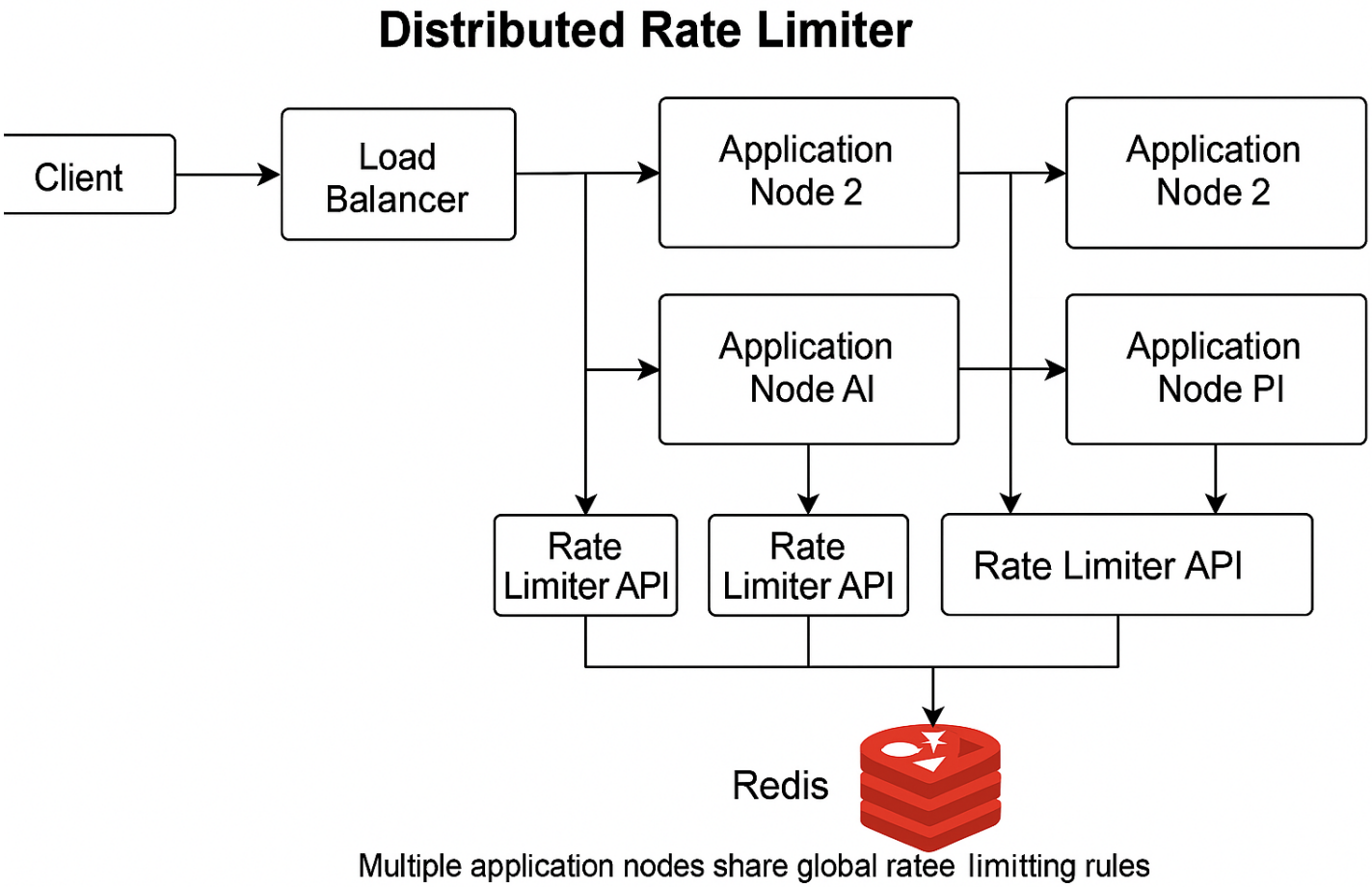
Distributed Rate Limiter: Visual Explanation



In a distributed rate limiter setup, multiple service nodes (Service A, B, C) behind a load balancer share the same Redis backend to enforce rate limits. This ensures that a client can't bypass limits by hitting different instances. Redis acts as the central counter with atomic operations and TTLs for expiring limits.

This architecture guarantees global consistency, even with horizontal scaling, and is crucial for APIs,

gateways, and infrastructure-level services.