Chapter 9 (Hindi Version)

Load Balancing Basics

Phase 1 · Fundamentals · Estimated time: 1.5 hours

1. Concept From First Principles

English:

One server can only handle so many requests before it runs out of CPU, memory, or connections. The fix is to run multiple copies of your application -- but then something needs to decide which server handles each incoming request. That something is a load balancer.

Hindi:

एक server एक सीमा तक ही requests handle कर सकता है, उसके बाद CPU, memory, या connections खत्म हो जाते हैं। इसका हल है अपने application की कई copies चलाना -- लेकिन फिर किसी को यह decide करना होगा कि हर आने वाली request किस server के पास जाएगी। वह कोई load balancer होता है।

Backend engineer analogy

English:

If you've ever run your Node/Laravel app behind Nginx with multiple PM2/PHP-FPM workers, Nginx was already acting as a mini load balancer, spreading requests across workers on the same machine.

Hindi:

अगर आपने कभी अपने Node/Laravel app को Nginx के पीछे कई PM2/PHP-FPM workers के साथ चलाया है, तो Nginx पहले से ही एक mini load balancer की तरह काम कर रहा था, requests को उसी machine के workers में बांटते हुए।

L4 vs L7 load balancing

English:

Layer 4 (L4) load balancers operate at the transport layer -- they route based only on IP address and port, without looking inside the actual HTTP request. Layer 7 (L7) load balancers inspect the actual HTTP request and can make smarter routing decisions.

Hindi:

Layer 4 (L4) load balancers transport layer पर काम करते हैं -- वे सिर्फ IP address और port के आधार पर route करते हैं, असली HTTP request के अंदर देखे बिना। Layer 7 (L7) load balancers असली HTTP request को देखते हैं और ज़्यादा smart routing decisions ले सकते हैं।

Important Words:

- **Route (verb)** – रूट करना / रास्ता तय करना

English:

Most production systems use both: an L4 balancer at the outer edge for raw throughput, with L7 balancers or an API gateway behind it for intelligent routing.

Hindi:

ज़्यादातर production systems दोनों use करते हैं: बाहर की तरफ raw throughput के लिए एक L4 balancer, और उसके पीछे intelligent routing के लिए L7 balancers या एक API gateway।

Load balancing algorithms

English:

- Round Robin: requests are distributed to servers in rotating order.
- Least Connections: send the next request to whichever server currently has the fewest active connections.
- Weighted variants: give some servers a higher share of traffic.
- IP Hash / Consistent Hashing: route the same client consistently to the same server.

Hindi:

- Round Robin: requests को बारी-बारी क्रम में servers पर भेजा जाता है।
- Least Connections: अगली request उस server को भेजी जाती है जसिके पास सबसे कम active connections हों।
- Weighted variants: कुछ servers को traffic का ज़्यादा हस्सि दयिा जाता है।
- IP Hash / Consistent Hashing: एक ही client को हमेशा एक ही server पर भेजा जाता है।

English:

A load balancer continuously pings each backend server and stops sending traffic to any server that fails to respond correctly, automatically routing around failures without human intervention.

Hindi:

एक load balancer लगातार हर backend server को ping करता रहता है और जो server सही जवाब नहीं देता उसे traffic भेजना बंद कर देता है, बनिा कसिी इंसानी दखल के अपने आप failures के आस-पास से रास्ता नकिाल लेता है।

Important Words:

- **Intervention** – हस्तक्षेप / दखल

English:

One nuance worth flagging: sticky sessions can look like a convenient shortcut, but they quietly undermine the benefits of horizontal scaling -- if that one server goes down, every client stuck to it loses their session.

Hindi:

एक बारीक बात: sticky sessions एक आसान shortcut जैसी लग सकती हैं, लेकनि यह चुपके से horizontal scaling के फायदों को कमज़ोर कर देती हैं -- अगर वह एक server down हो जाए, तो उससे जुड़े हर client का session खो जाता है।

Important Words:

- **Undermine** – कमज़ोर करना

English:

It's worth being precise about what a load balancer does not solve: it distributes traffic across existing capacity, but it doesn't create new capacity. Load balancing and auto-scaling are complementary, not interchangeable.

Hindi:

यह साफ समझना ज़रूरी है कि load balancer क्या नहीं करता: यह मौजूदा capacity में traffic बांटता है, नई capacity नहीं बनाता। Load balancing और auto-scaling एक-दूसरे के पूरक हैं, एक-दूसरे का विकल्प नहीं।

Important Words:

- **Complementary** – पूरक / एक-दूसरे को पूरा करने वाला

2. Real-World Example / Case Study

English:

During a Flipkart Big Billion Days sale, the load balancer layer is what lets Flipkart absorb a traffic spike by simply adding more application server instances -- without changing a single line of app code.

Hindi:

Flipkart की Big Billion Days sale के दौरान, load balancer layer ही Flipkart को traffic बढ़ने पर संभालने देती है, बस ज़्यादा application server instances जोड़कर -- बिना app code की एक भी लाइन बदले।

Important Words:

- **Spike (traffic)** – अचानक बढ़ोतरी

English:

Careem's API gateway routes ride-booking, payment, and food-delivery requests to entirely different backend service pools based on the URL path -- a textbook L7 routing decision.

Hindi:

Careem का API gateway ride-booking, payment, और food-delivery requests को URL path के आधार पर बिल्कुल अलग-अलग backend service pools पर भेजता है -- यह L7 routing का एक textbook उदाहरण है।

English:

A second example: Noon's search service benefits particularly from least-connections load balancing since search queries vary widely in how long they take to process.

Hindi:

एक और उदाहरण: Noon की search service को least-connections load balancing से खास फायदा होता है क्योंकि search queries को process होने में लगने वाला समय काफी अलग-अलग होता है।

3. Diagram (English labels kept as-is)

```
                        [ Load Balancer ]
                          health checks -->
                        /        |        \
                       v         v         v
                 [ Server 1 ] [ Server 2 ] [ Server 3 ]
                  (healthy)    (healthy)   (FAILED - removed)

L7 example - routing by path:
  /api/payments/*  ---> Payment Service pool
  /api/search/*    ---> Search Service pool
```

4. Trade-offs Table (Reference Table – English)

Layer / Algorithm	Strength	Weakness
L4	Very fast, low overhead	Can't route on URL/headers
L7	Smart routing on path/headers/cookies	More CPU/latency cost
Round Robin	Simple, predictable	Ignores actual server load
Least Connections	Adapts to uneven request durations	Slightly more overhead

5. Common Interview Questions

Q1: "Where would you place a load balancer, and why?"

English:

Model approach: Place it in front of any tier with multiple stateless instances. Explain it removes a single point of failure and enables horizontal scaling without client-visible changes.

Hindi:

Model approach: इसे किसी भी tier के आगे रखें जिसके कई stateless instances हों। समझाएं कि यह single point of failure हटाता है और client को कुछ बदले बिना horizontal scaling करने देता है।

Q2: "How does the load balancer itself avoid becoming a single point of failure?"

English:

Model approach: Explain that production load balancers are typically deployed in redundant pairs or clusters with a floating IP or DNS-based failover.

Hindi:

Model approach: समझाएं कि production load balancers आमतौर पर redundant pairs या clusters में deploy होते हैं, साथ में floating IP या DNS-based failover होता है।

Q3: "Why is sticky session routing generally discouraged?"

English:

Model approach: It ties a client to one server, reintroducing single-point-of-failure and uneven-load problems. The preferred fix is making the service stateless.

Hindi:

Model approach: यह client को एक ही server से बांध देता है, जिससे single-point-of-failure और असमान load जैसी समस्याएं फिर से आ जाती हैं। बेहतर हल है service को stateless बनाना।

6. Practice Exercises

English:

- Explain the difference between L4 and L7 load balancing to a junior engineer.
- For an app with separate search, checkout, and recommendation services, sketch an L7 routing table.
- Describe what happens, step by step, when one of three backend servers suddenly crashes.

Hindi:

- L4 और L7 load balancing का फर्क किसी junior engineer को समझाएं।
- अलग-अलग search, checkout, और recommendation services वाले app के लिए एक L7 routing table बनाएं।
- बताएं कि जब तीन backend servers में से एक अचानक crash हो जाए, तो step-by-step क्या होता है।

7. Curated YouTube Videos (Reference – English)

Language	Video & Channel	Why it helps
English	"Load Balancers L4 vs L7 Explained"	Directly covers this chapter's core distinction.
English	"Load Balancer in System Design Interviews"	Frames it for interview answers.
Hindi	"System Design: Load Balancer Explained"	Hindi walkthrough of core concepts.
Hindi	"Complete System Design Roadmap 2025" — CodeHelp	Covers load balancing in a Hindi roadmap.

8. Key Takeaways

English:

- A load balancer distributes incoming traffic across multiple servers so no single one is overwhelmed.
- L4 routes on IP/port only (fast); L7 routes on request content (smart, costlier).
- Common algorithms: round robin, least connections, weighted variants, consistent hashing.
- Health checks let the load balancer automatically stop sending traffic to failed servers.
- Load balancers themselves must be deployed redundantly to avoid becoming a single point of failure.
- Load balancing decouples 'how much traffic we can handle' from 'how the application is written.'

Hindi:

- एक load balancer आने वाले traffic को कई servers में बांटता है ताकि कोई एक server ओवरलोड ना हो।
- L4 सिर्फ IP/port पर route करता है (तेज़); L7 request की content पर route करता है (smart, महंगा)।
- आम algorithms: round robin, least connections, weighted variants, consistent hashing।
- Health checks load balancer को failed servers पर traffic भेजना अपने आप बंद करने देते हैं।
- Load balancers को खुद भी redundant तरीके से deploy करना ज़रूरी है ताकि वे single point of failure ना बनें।
- Load balancing 'हम कितना traffic संभाल सकते हैं' को 'application कैसे लिखा गया है' से अलग कर देता है।

General Words Glossary

A consolidated list of the important English words used throughout this chapter, with Hindi meanings and simple explanations — useful for revising vocabulary after finishing the chapter.

English Word	Hindi Meaning	Simple Explanation
Complementary	पूरक / एक-दूसरे को पूरा करने वाला	Working together to complete each other
Intervention	हस्तक्षेप / दखल	Someone stepping in to fix or change something
Route (verb)	रूट करना / रास्ता तय करना	To direct traffic to a destination
Spike (traffic)	अचानक बढ़ोतरी	A sudden sharp increase
Undermine	कमज़ोर करना	To weaken something gradually