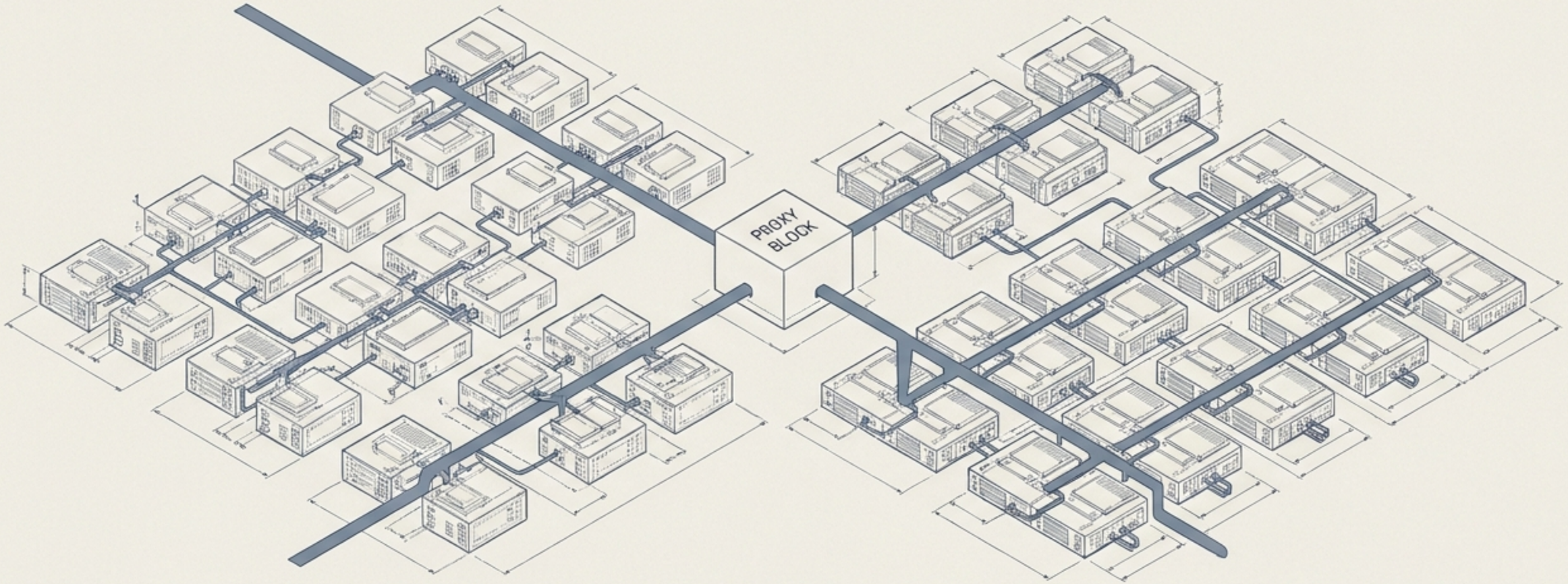
WEB ARCHITECTURE — S02E04

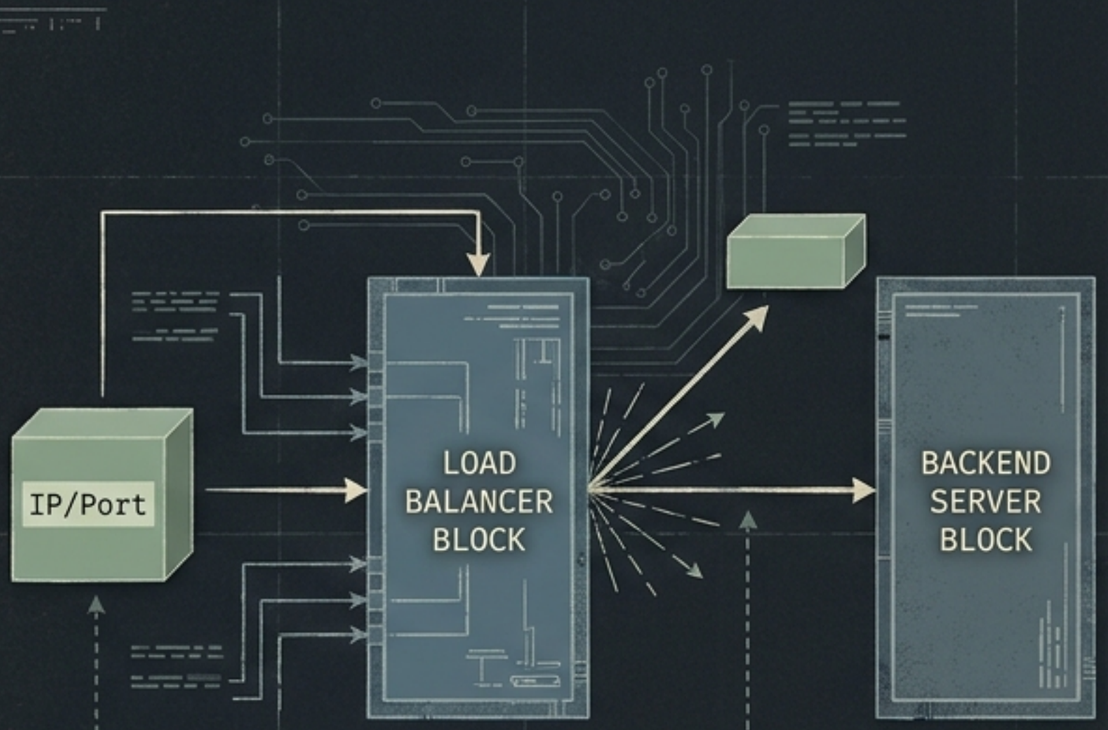
Reverse Proxies, L4 vs. L7 Load Balancing, HTTP/2 & QUIC Multiplexing, and Health-Check Engineering



ENV: PRODUCTION | **AUDIENCE:** SRE, BACKEND, PLATFORM | **STATUS:** DEEP_DIVE

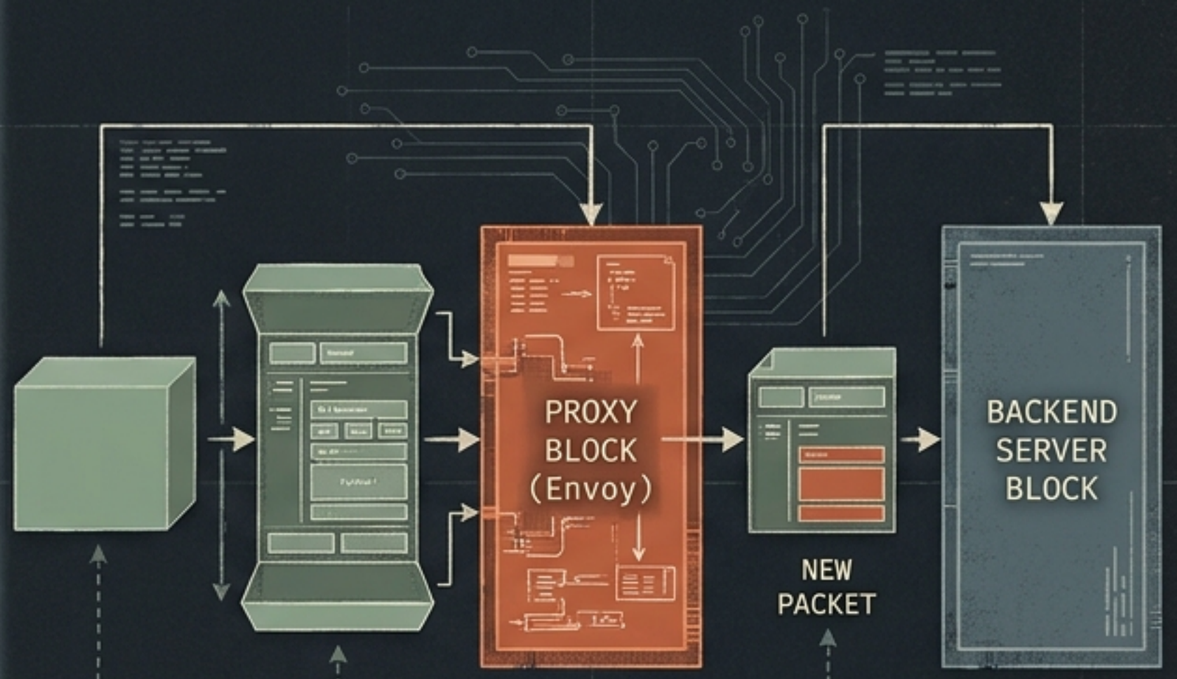
Gemini Notebook

Layer 4 (Transport)



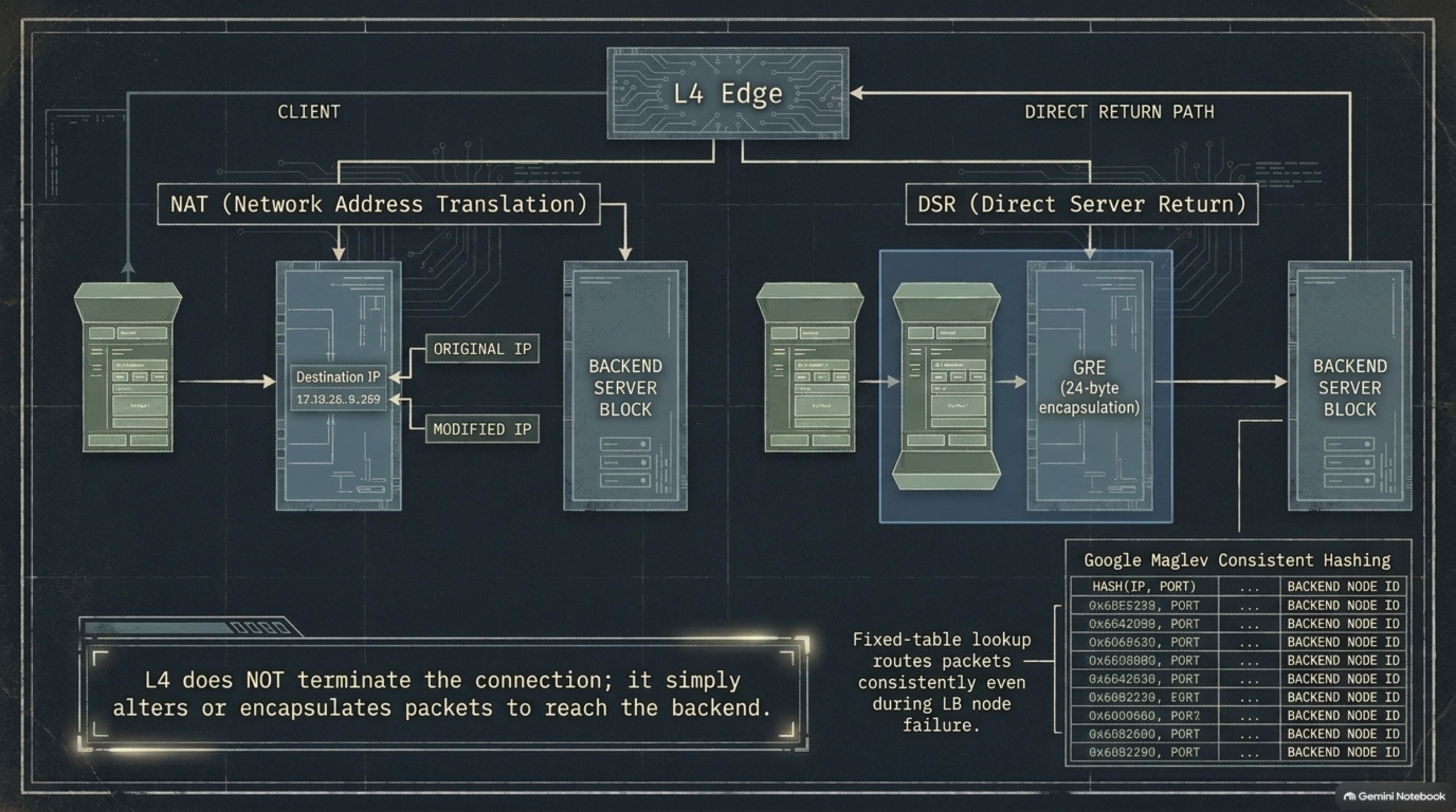
Ignorant but blindingly fast.
Forwards packets based on IP/Port.

Layer 7 (Application)



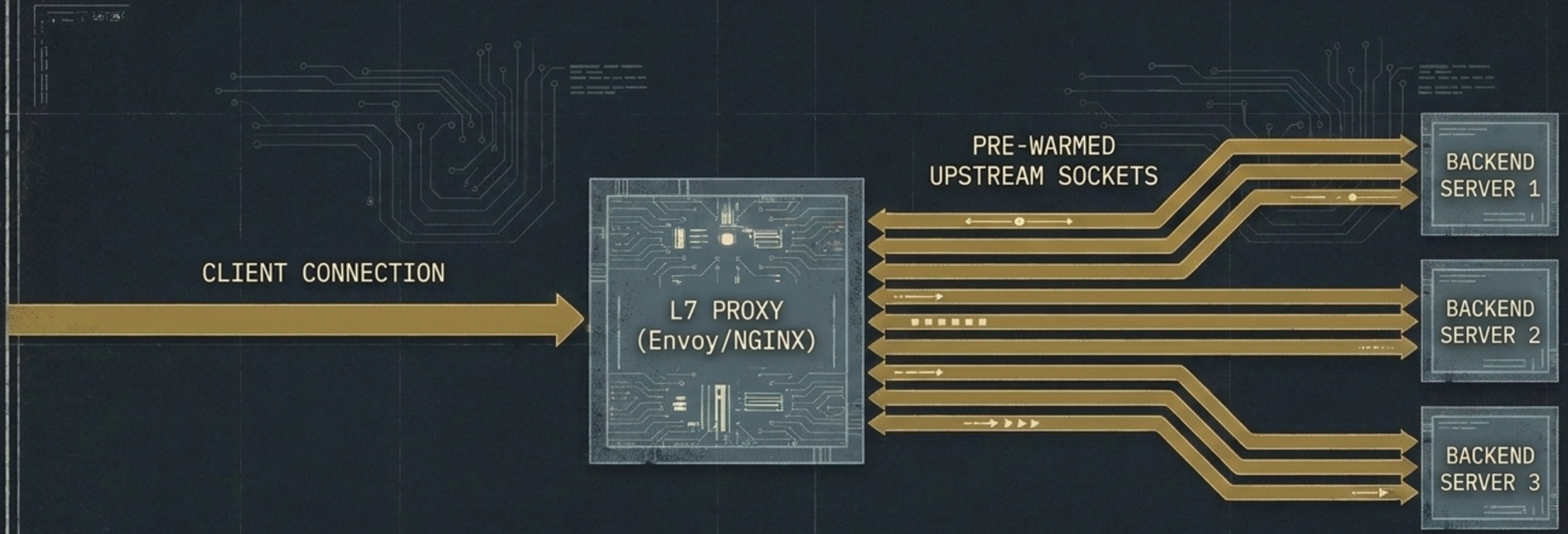
Intelligent but computationally expensive.
Reads application semantics.

Gemini Notebook



Gemini Notebook

The Reverse Proxy Disconnect



Envoy, NGINX, AWS ALB.

Terminated downstream client sockets ≠ Pooled upstream backend sockets.

Enables protocol translation, caching, and request-level routing.

Gemini Notebook

HTTP/1.1 & The Connection Exhaustion Limit



Browser Max Connection Limit

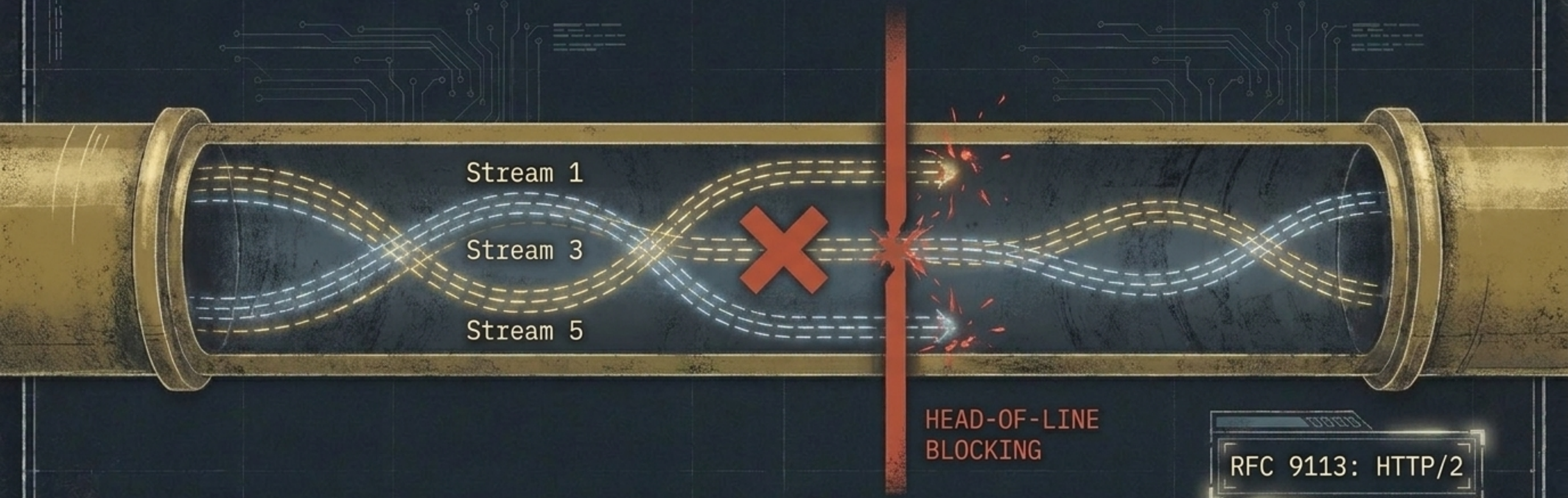
One request per connection.

High latency, massive TCP handshake overhead.

Application-layer head-of-line blocking.

Gemini Notebook

HTTP/2: Binary Framing & Head-of-Line Blocking



RFC 9113: HTTP/2

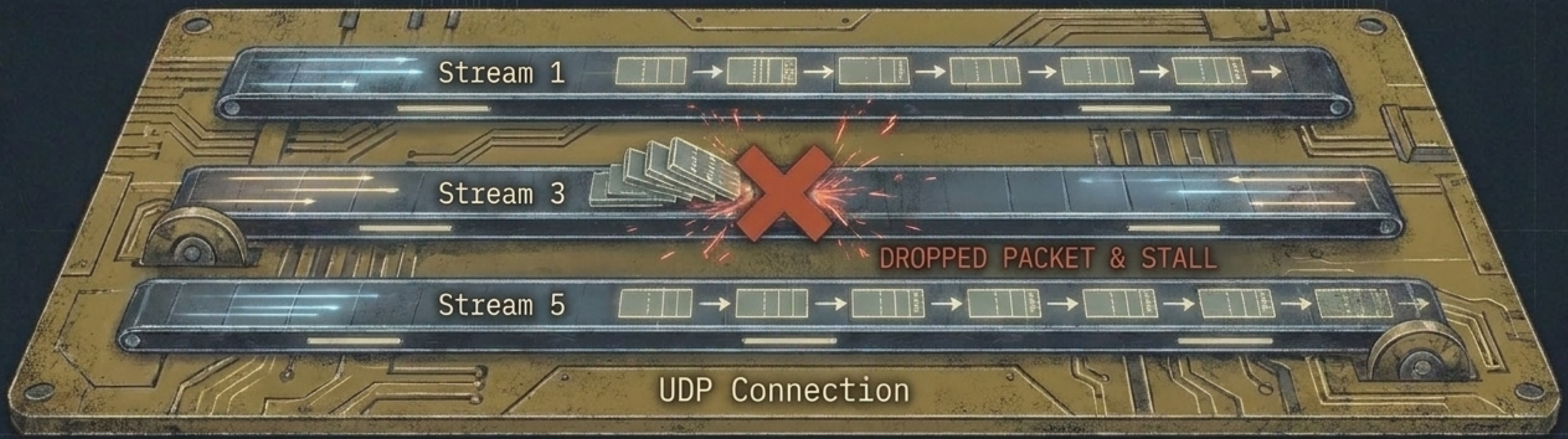
While HTTP/2 solved application-layer blocking via stream multiplexing, it pushed the vulnerability down to the OS kernel TCP layer. One lost packet stalls the entire connection.

Gemini Notebook

HTTP/3 & QUIC: Independence at the Transport Layer

RFC 9000 / 9114

Integrated TLS 1.3



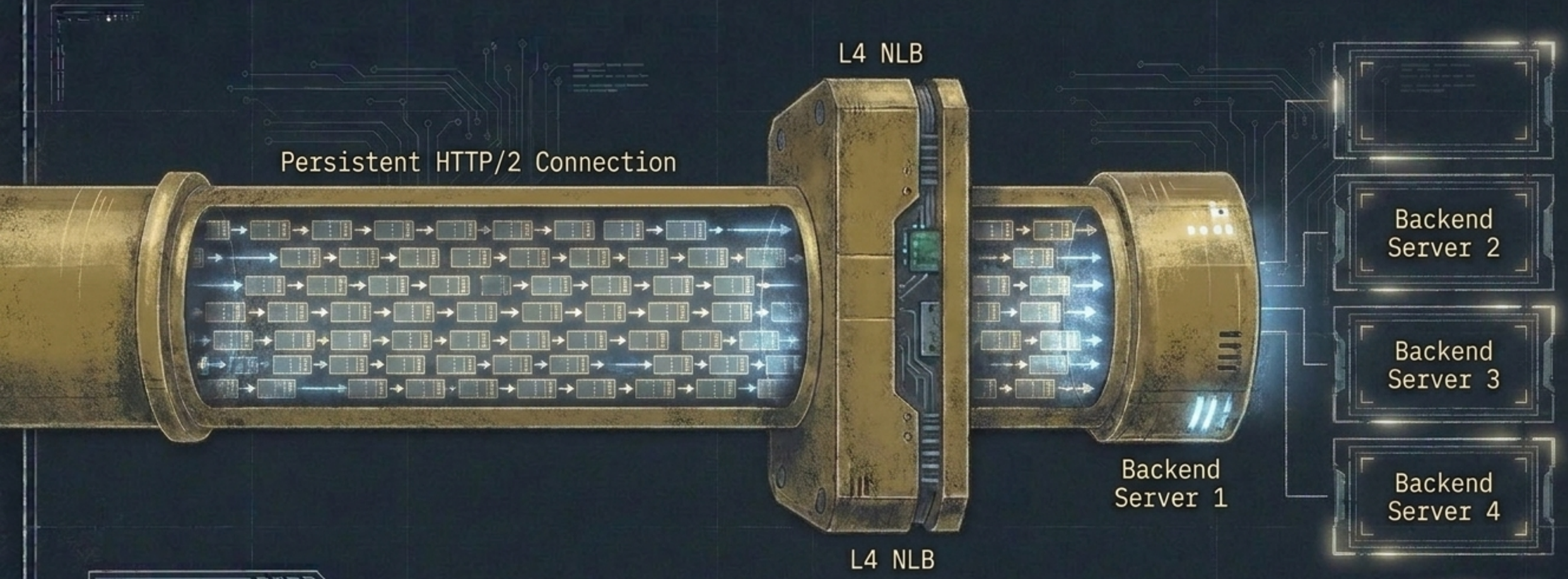
Independent stream loss recovery over UDP.

Zero RTT connection establishment.

Connection IDs survive client IP/port changes.

Gemini Notebook

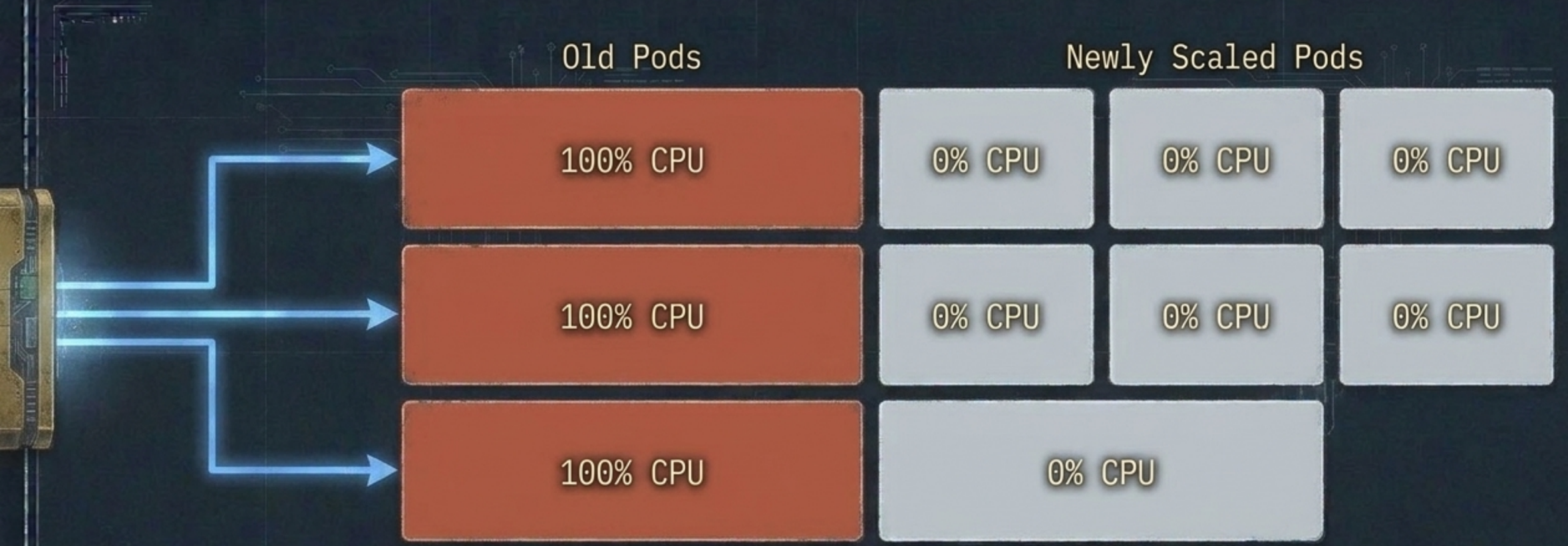
The Trap: Mixing L4 Routing with Persistent Protocols



Because L4 load balancers route CONNECTIONS, and HTTP/2/gRPC multiplexes all requests onto a SINGLE persistent connection, the LB is rendered useless after the initial handshake.

Gemini Notebook

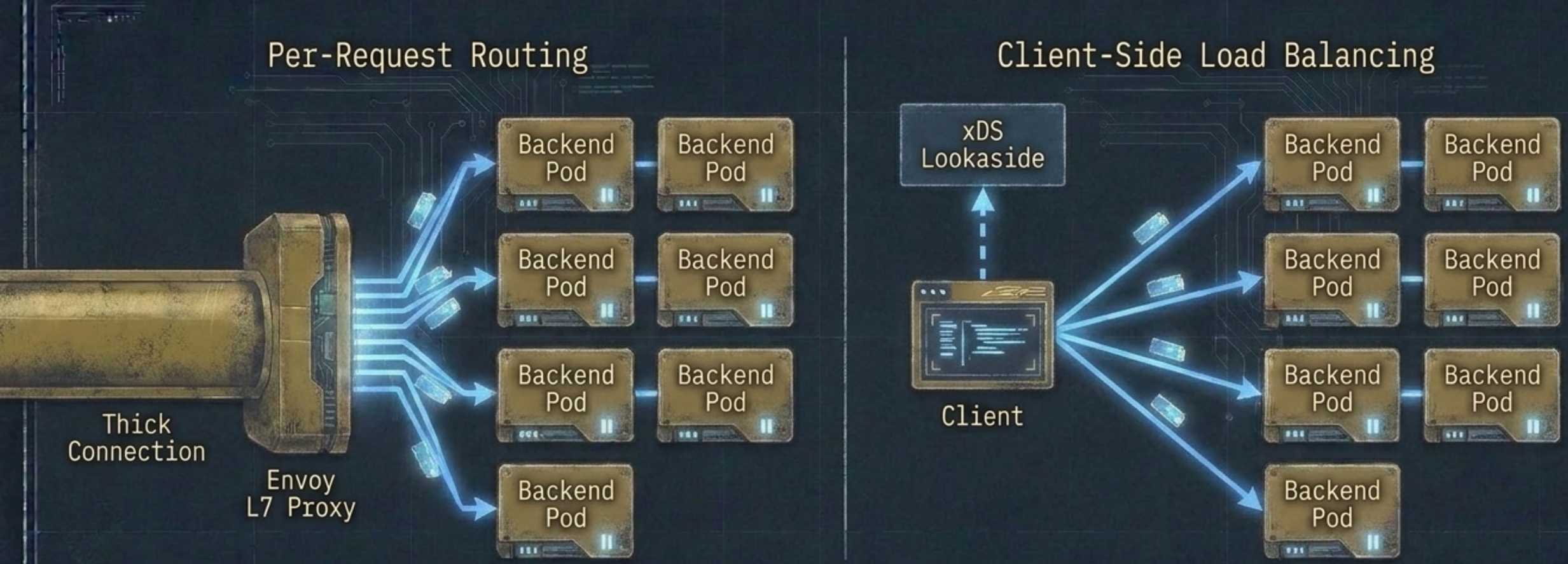
The Scale-Out Hotspot



Connection-level L4 balancing pins persistent gRPC/HTTP2 traffic to old pods. During scale-out events, new pods receive zero traffic because clients never establish new TCP connections.

Gemini Notebook

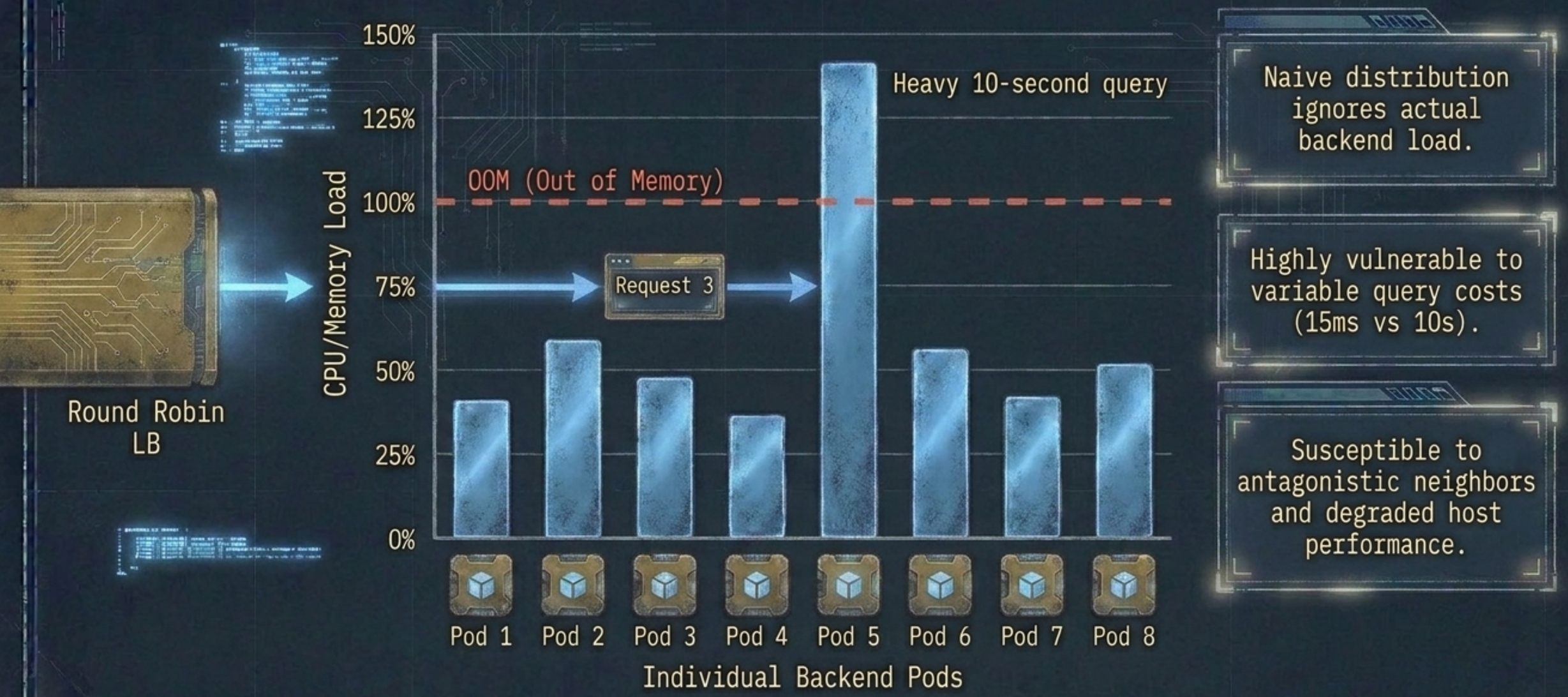
The Resolution: L7 Routing or Client-Side xDS



Shift from connection-level routing to request-level routing.

Gemini Notebook

The Pitfalls of Simple Round Robin



Naive distribution ignores actual backend load.

Highly vulnerable to variable query costs (15ms vs 10s).

Susceptible to antagonistic neighbors and degraded host performance.

Gemini Notebook

Least Request & Power of Two Random Choices (P2C)



Avoids O(N) scanning overhead of strict Least Request.

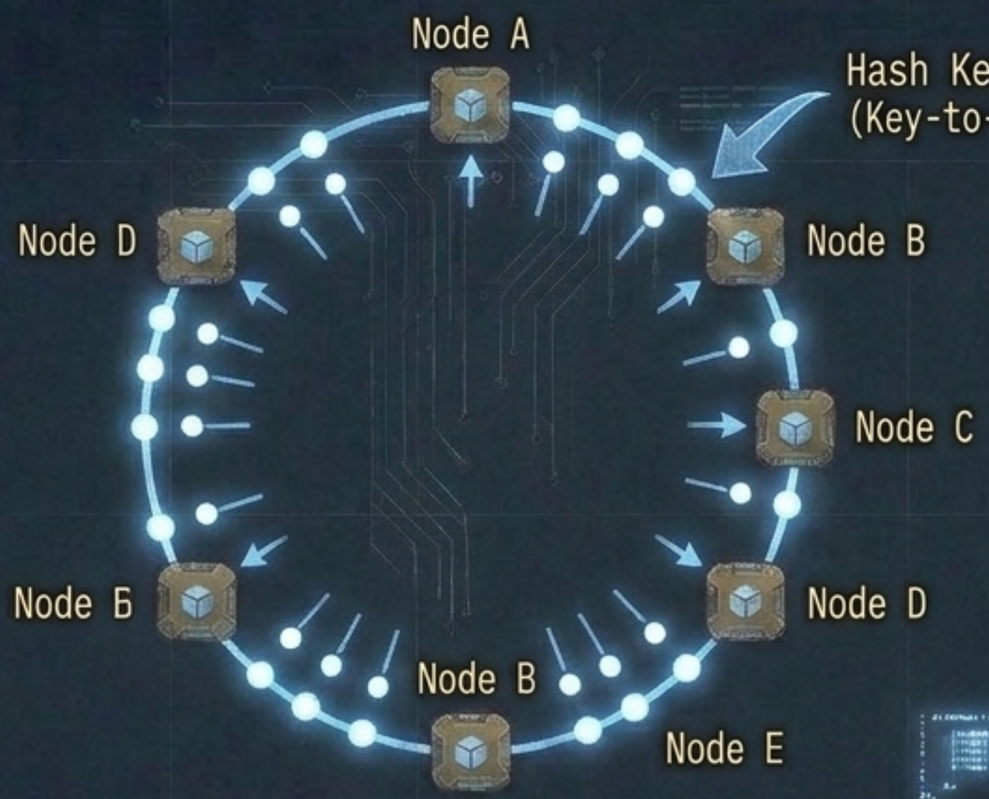
Prevents thundering herd behavior when telemetry is slightly stale.

Weights adjust dynamically:

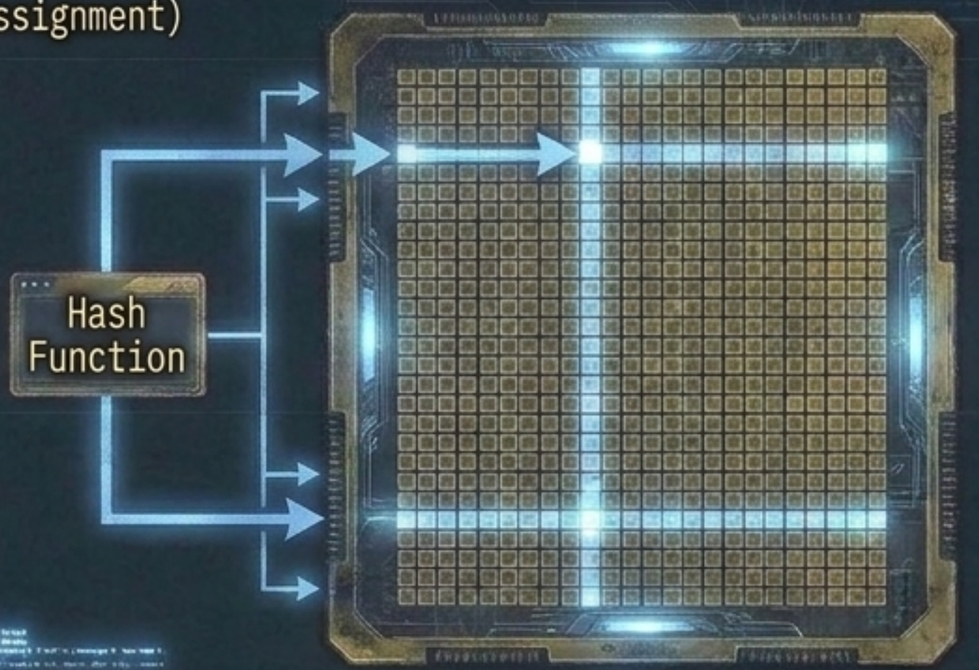
$$\text{weight} / (\text{active_requests} + 1)^{\text{active_request_bias}}$$

Gemini Notebook

Consistent Hashing: Ring Hash vs. Maglev



Ring Hash (Ketama) - Key Mapping



Maglev Lookup Table - Matrix Mapping

Locality-aware routing for cache-affinity.

Maglev provides substantially faster host selection (~5x) and table build times (~10x) over Ring Hash.

Minimizes disruption and key reassignment when upstream hosts change.

Gemini Notebook

THE ALGORITHM DECISION MATRIX

ALGORITHM	COMPUTE OVERHEAD	LOAD DISTRIBUTION	STATEFULNESS	CHURN RESILIENCE
ROUND ROBIN	VERY LOW	POOR	STATELESS	N/A
P2C (LEAST REQUEST)	LOW	EXCELLENT	STATELESS	N/A
RING HASH	MEDIUM	GOOD	STATEFUL (CACHE AFFINITY)	MODERATE
MAGLEV	LOW	GOOD	STATEFUL (CACHE AFFINITY)	HIGH

Gemini Notebook

Upstream Connection Pooling



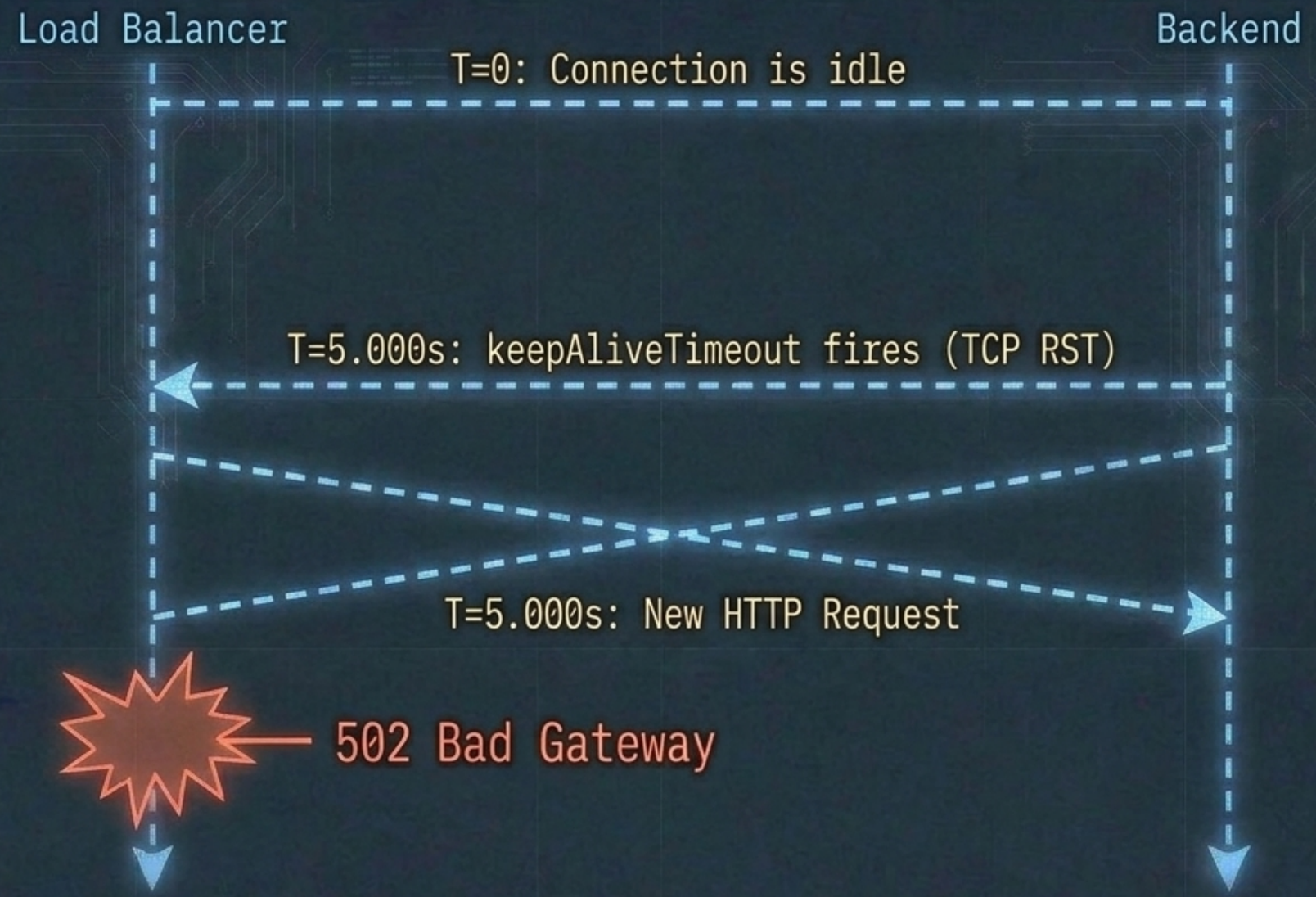
Amortizing the computational and latency costs of TCP/TLS handshakes.

Tuning max_requests_per_connection limits connection aging.

Automatic protocol selection via ALPN.

Gemini Notebook

The 502 Bad Gateway Race Condition



Gemini Notebook

Eliminating the 502 Race

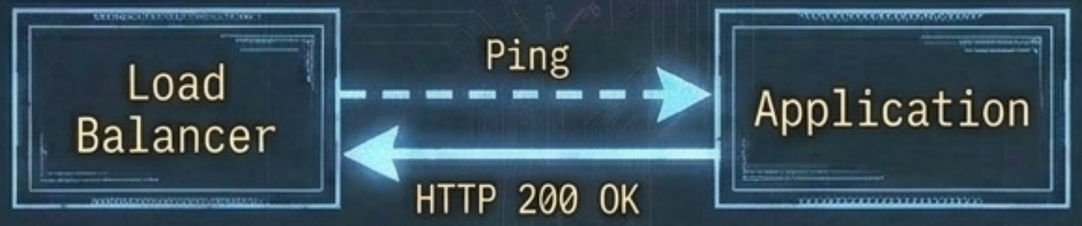
Backend keepAliveTimeout > **Load Balancer idleTimeout**

The Golden Rule of Timeout Tuning: Always configure the downstream component (the backend) to hold the connection open slightly longer than the upstream component (the proxy). This ensures the proxy always closes idle connections predictably.

Gemini Notebook

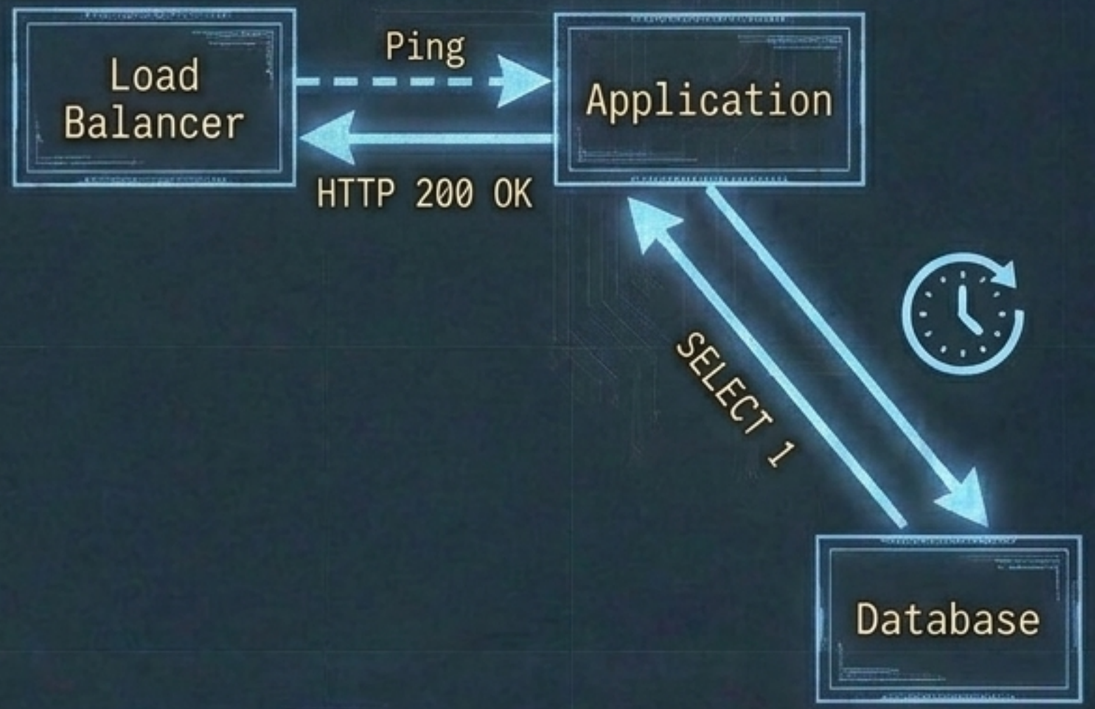
Shallow vs. Deep Health Checks

Shallow Check



Shallow (L4/Process):
Am I listening?
(Fast, cheap)

Deep Check



Deep (L7/Dependency):
Can I reach the database?
(Slow, dependency-bound)

Gemini Notebook

The Eviction Cascade (The Danger of Deep Checks)

1. Database experiences minor latency spike.

2. Deep health checks time out.

3. Load balancer evicts 10% of pods.

4. Remaining 90% of pods receive MORE traffic.

5. Database gets hammered harder, slows down more.

6. More health checks fail.

5. Forced or health eviction.



⚠ Result: 100% fleet eviction (Blackout) ⚠

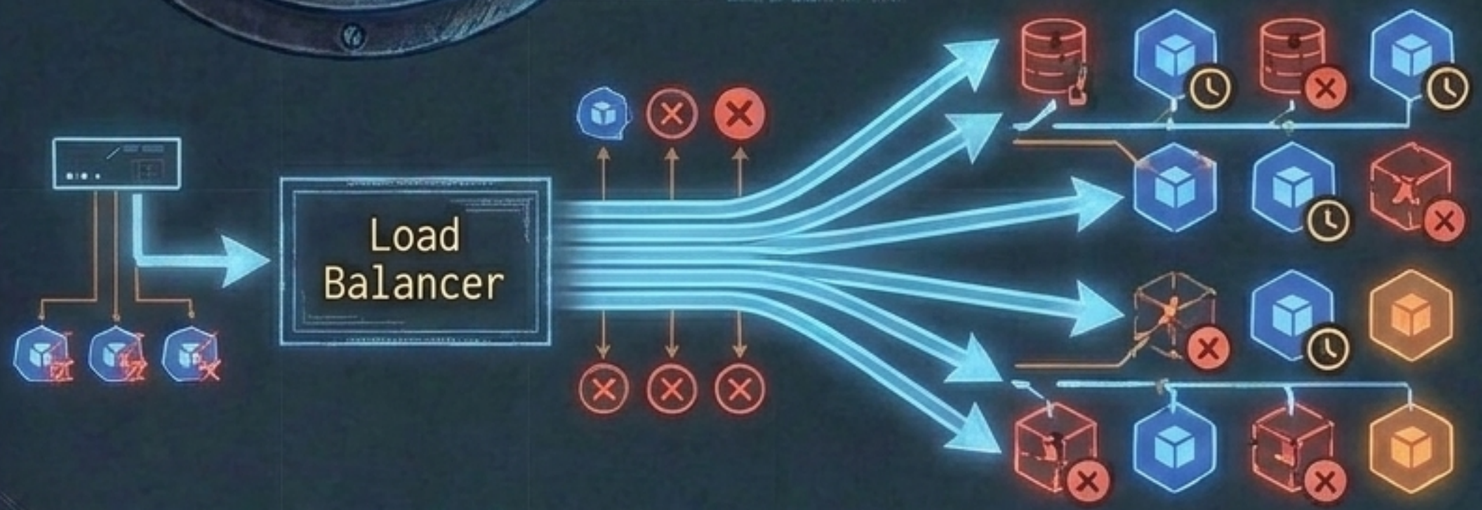
How a partial downstream brownout turns into a self-inflicted 100% outage.

Gemini Notebook

Fail-Open Safety & Panic Thresholds



PANIC THRESHOLD TRIGGERED



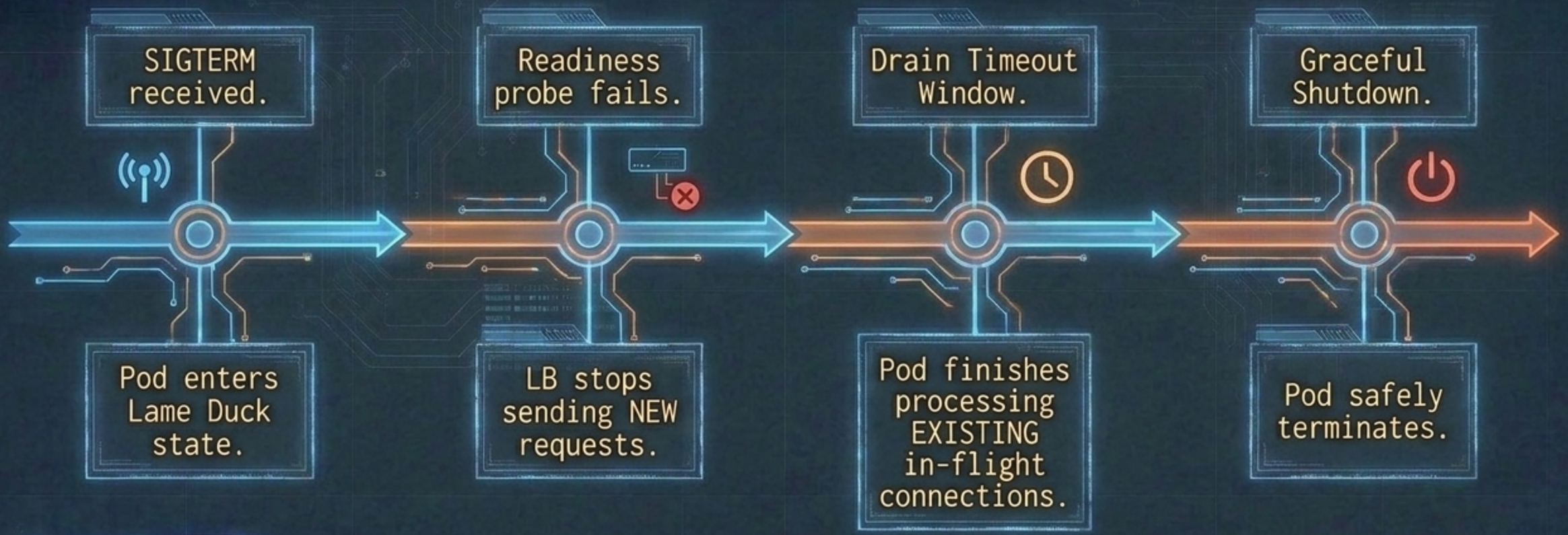
Panel 1:
If the number of healthy hosts drops below a critical threshold (e.g., 50%), assume the health-checker itself is wrong.

Panel 2:
The load balancer fails-open, routing traffic to ALL nodes (healthy or not).

Panel 3:
Prevents total fleet eviction during transient network partitions.

Gemini Notebook

Graceful Connection Draining



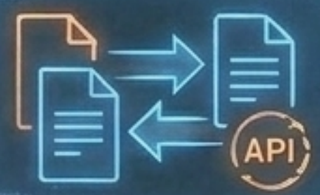
Handling scale-downs, code pushes, and spot-instance terminations with zero dropped packets.

Gemini Notebook

Identifying Your Workload Archetype



High-TPS TCP
(Streaming/IoT data)



Standard REST
(Public-facing web APIs)



WebSockets
(Real-time, persistent)

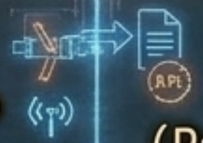


Internal gRPC Meshes
(East-West traffic)

Architectural choices must map strictly to the protocol and persistency profile of the workload.

Gemini Notebook

The Ultimate Infrastructure Decision Matrix

Solution	Best For	Terminating TLS?	Routing Intel	Latency Overhead
L4 NLB	High-TPS TCP / UDP (Streaming/IoT) 	No (Pass-through)	Low (IP/Port only)	Ultra-Low
L7 ALB / Edge Proxy	Standard REST / WebSockets (Public Ingress)	Yes	High (Paths, Headers)	Moderate
Service Mesh (Envoy/xDS)	Internal gRPC (East-West Microservices)	Yes (mTLS)	High (RPC methods)	Low (Sidecar/Lookaside)

Gemini Notebook

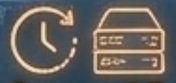
Key Engineering Takeaways



Don't mix L4 load balancing with multiplexed HTTP/2 or gRPC; use L7 or client-side xDS.



Adopt P2C (Power of Two Choices) for the best dynamic load balancing with the lowest overhead.



Tune timeouts: Backend keepAliveTimeout MUST strictly exceed proxy idleTimeout to prevent 502s.



Beware deep health checks: Rely on shallow checks combined with fail-open panic thresholds to prevent cascading fleet evictions.

Scale is not just about routing packets; it's about anticipating failure modes.

Gemini Notebook