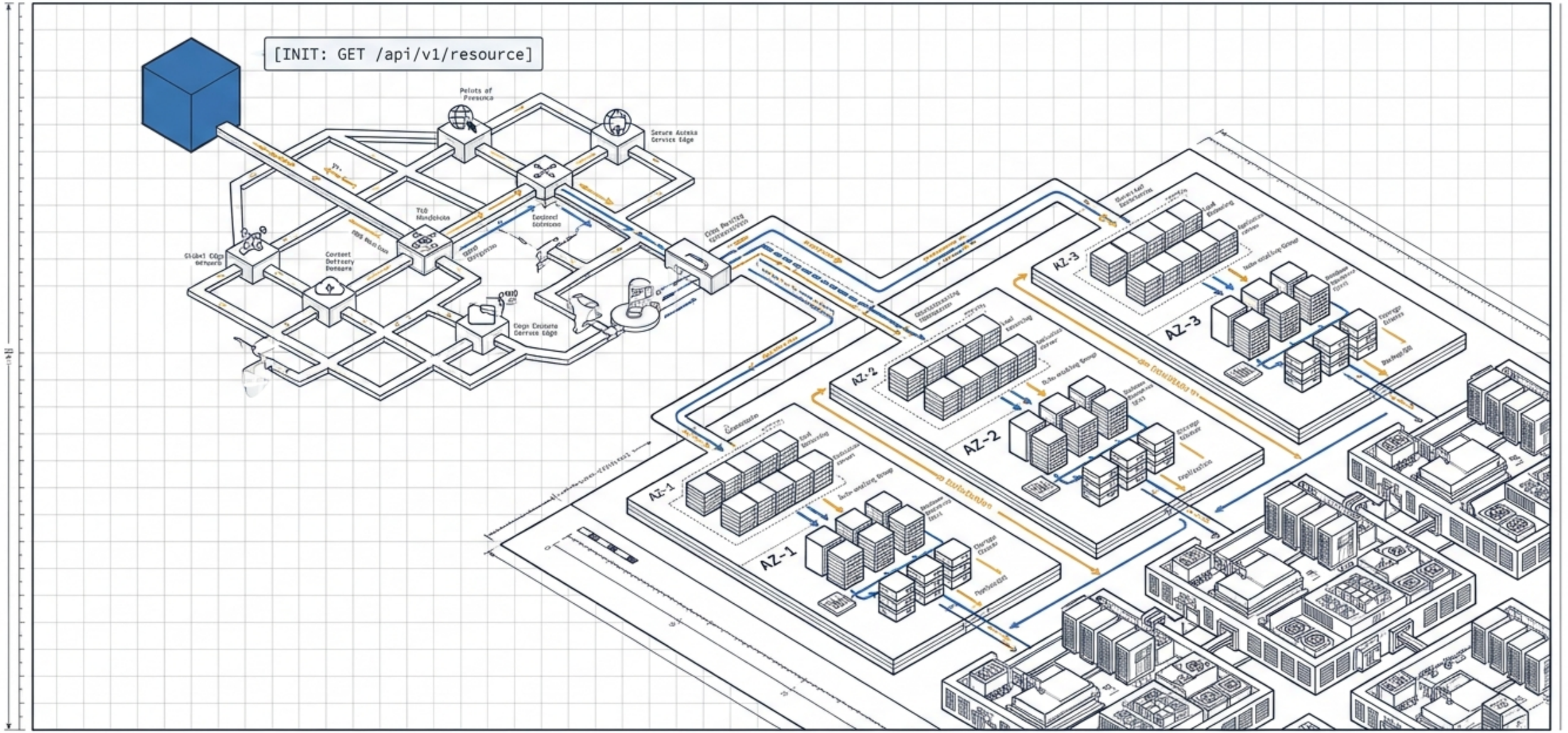
Web Architecture — S02E00: Life of a Production Request

An End-to-End Infrastructure Map & Resilience Blueprint



Gemini Notebook

THE PHYSICS AND MATH OF A

SINGLE-SERVER COLLAPSE

BLAST RADIUS DIAGRAM

CRITICAL FAILURE ANALYSIS

GEOGRAPHIC RTT



Signal propagation delay due to the speed of light and network infrastructure over global distances. Latency is unavoidable and cumulative.

VOLUMETRIC ATTACK SURFACE



A single IP absorbs all network attacks without a buffer. Saturation is immediate, leading to complete service denial.

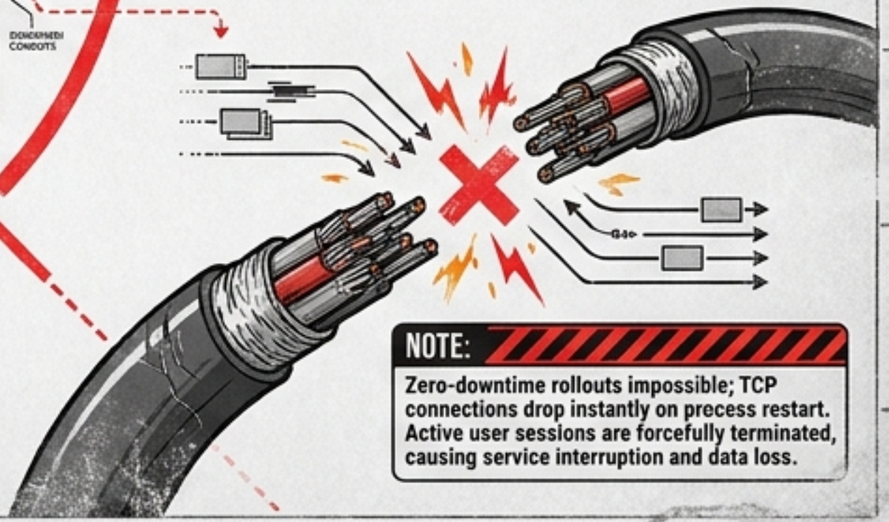
VERTICAL SCALING CEILINGS



Finite resources hit absolute performance walls. Exponential degradation and systemic failure occur past limits, resulting in process termination.



DEPLOY CONNECTION DROPS



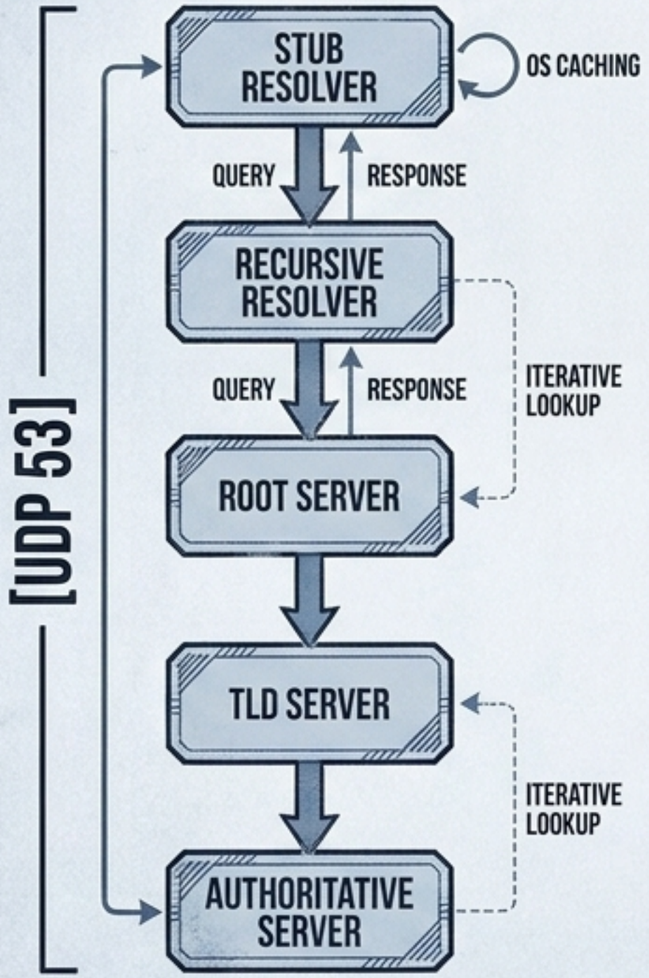
NOTE:

Zero-downtime rollouts impossible; TCP connections drop instantly on process restart. Active user sessions are forcefully terminated, causing service interruption and data loss.

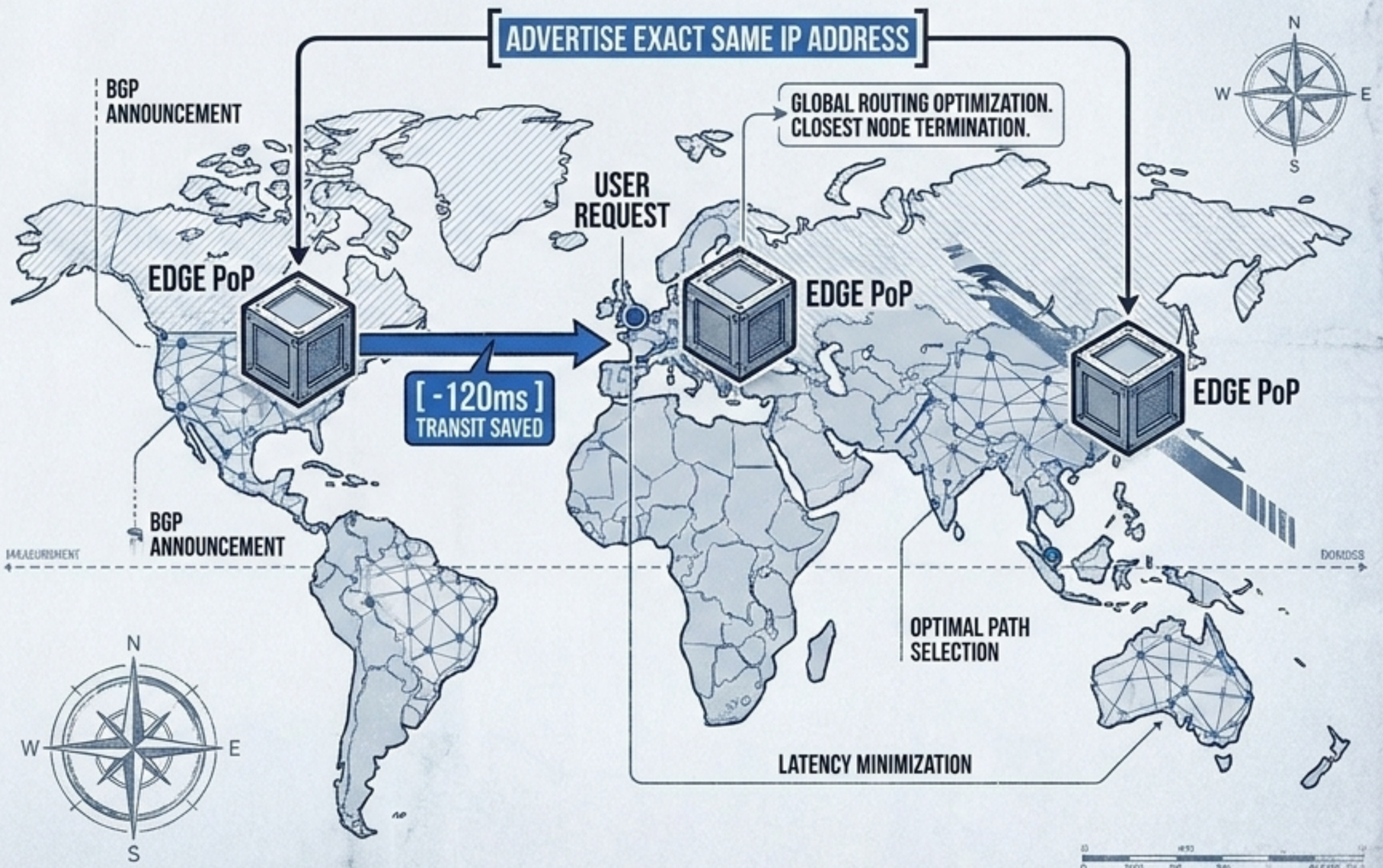
Gemini Notebook

THE OUTER SHIELD: DNS & BGP ANYCAST INGRESS

DNS RESOLUTION PATH



BGP ANYCAST ROUTING



Gemini Notebook

TERMINATING RISK AT THE EDGE (L7 WAF & CDN)

FILTER AND BOUNCE

MALICIOUS TRAFFIC

L7 WAF & DDoS

[Rate Limit / SQLi Blocked]

TLS 1.3 Termination

Origin compute cycles saved by terminating cryptographic handshakes at the edge.

RFC 9111 CDN Cache

[Cache-Hit Short-Circuit: <10ms]

LAYER STACK

ISOMETRIC FILTRATION PROCESS

EFFICIENCY GAIN

DYNAMIC TRAFFIC

Only legitimate, dynamic traffic proceeds to the Cloud Origin.

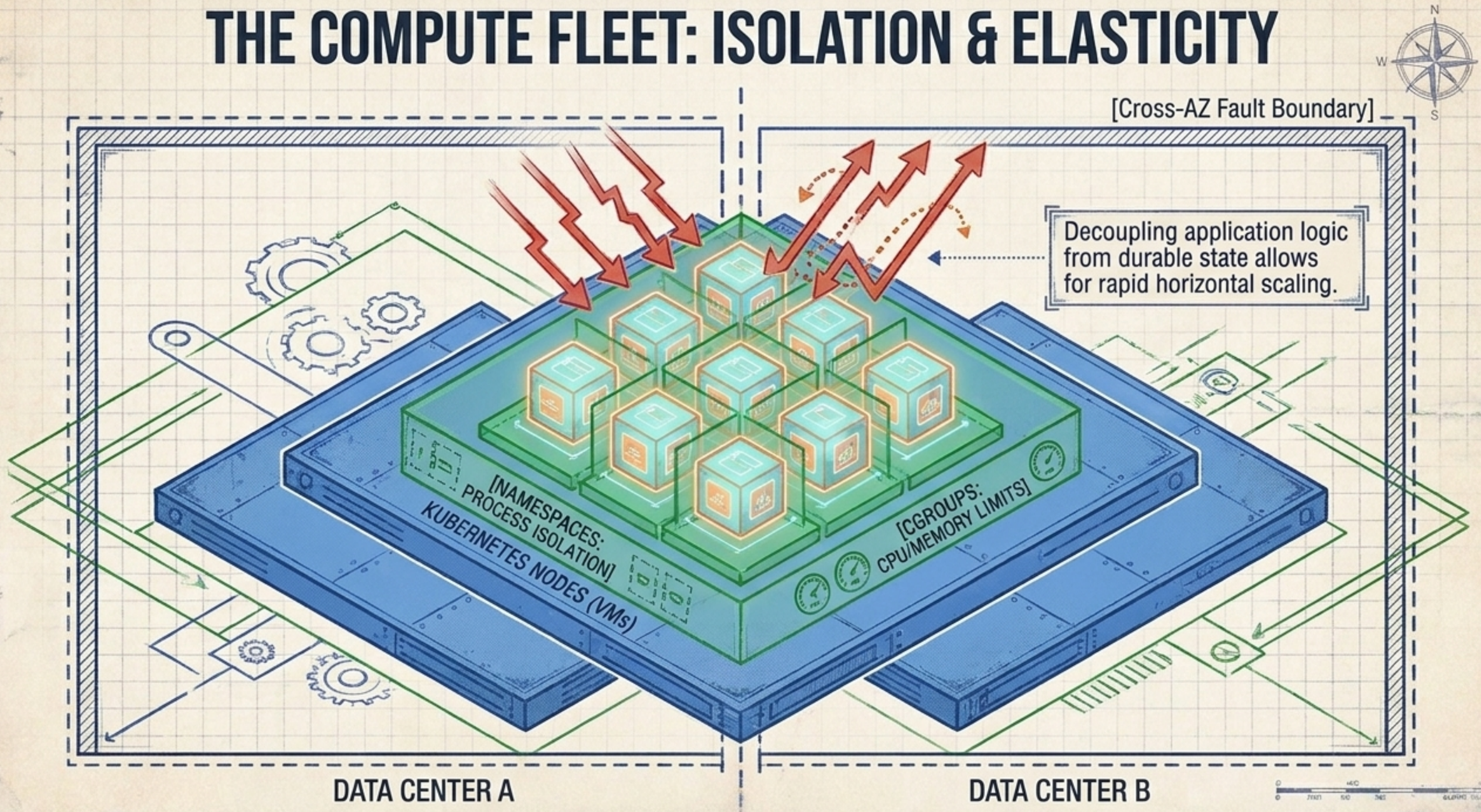
Gemini Notebook

DISAMBIGUATING THE DISTRIBUTION LAYER

FEATURE	CDN	L4 NETWORK LOAD BALANCER (NLB)	L7 APPLICATION LOAD BALANCER (ALB)	API GATEWAY
OSI LAYER	L7 (Edge)	L4 (Transport)	L7 (Application)	L7 (Application)
STATEFULNESS	Stateless & Cached	Stateless Packet-based (Consistent Hashing)	Stateful Connection-based	Stateful
PROTOCOL AWARENESS	HTTP/HTTPS	TCP/UDP	HTTP/gRPC/ WebSocket	Auth/Rate-limits/ Routing
LATENCY COST	<10ms	~1ms	~3-5ms	~10-20ms

Gemini Notebook

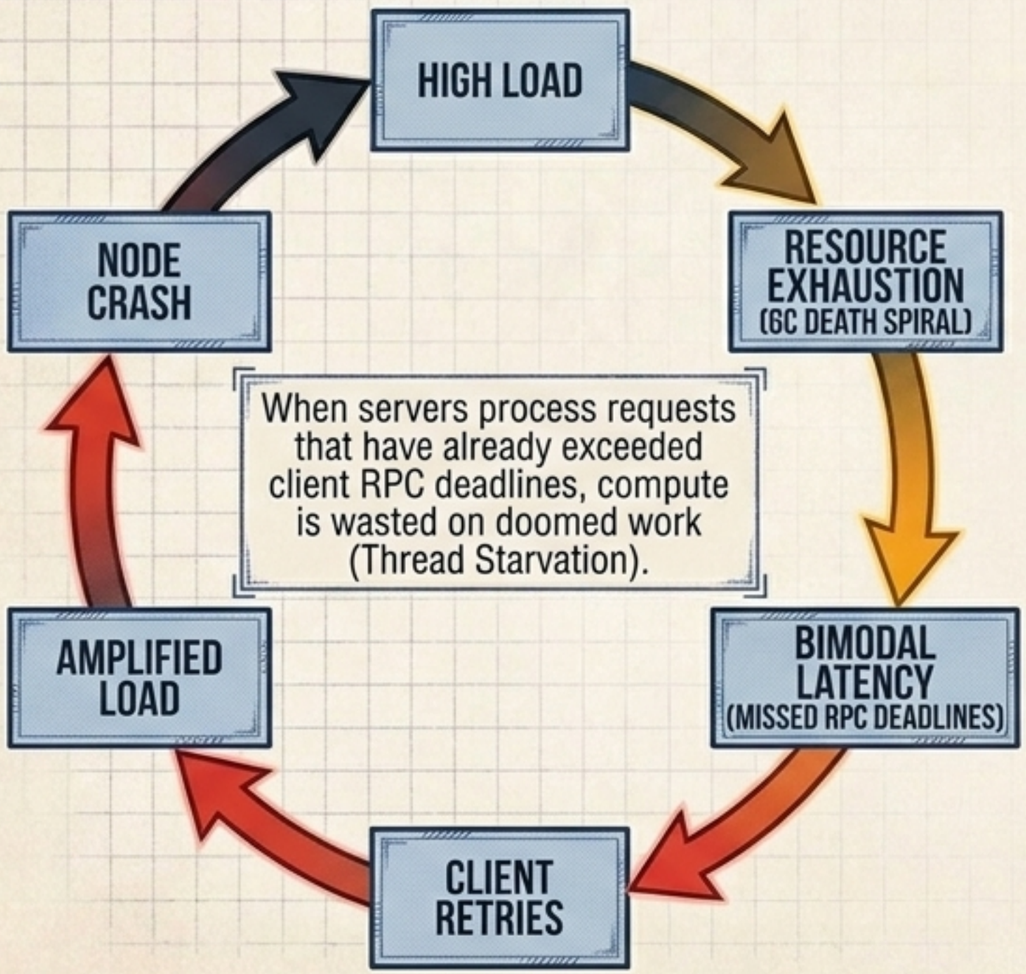
THE COMPUTE FLEET: ISOLATION & ELASTICITY



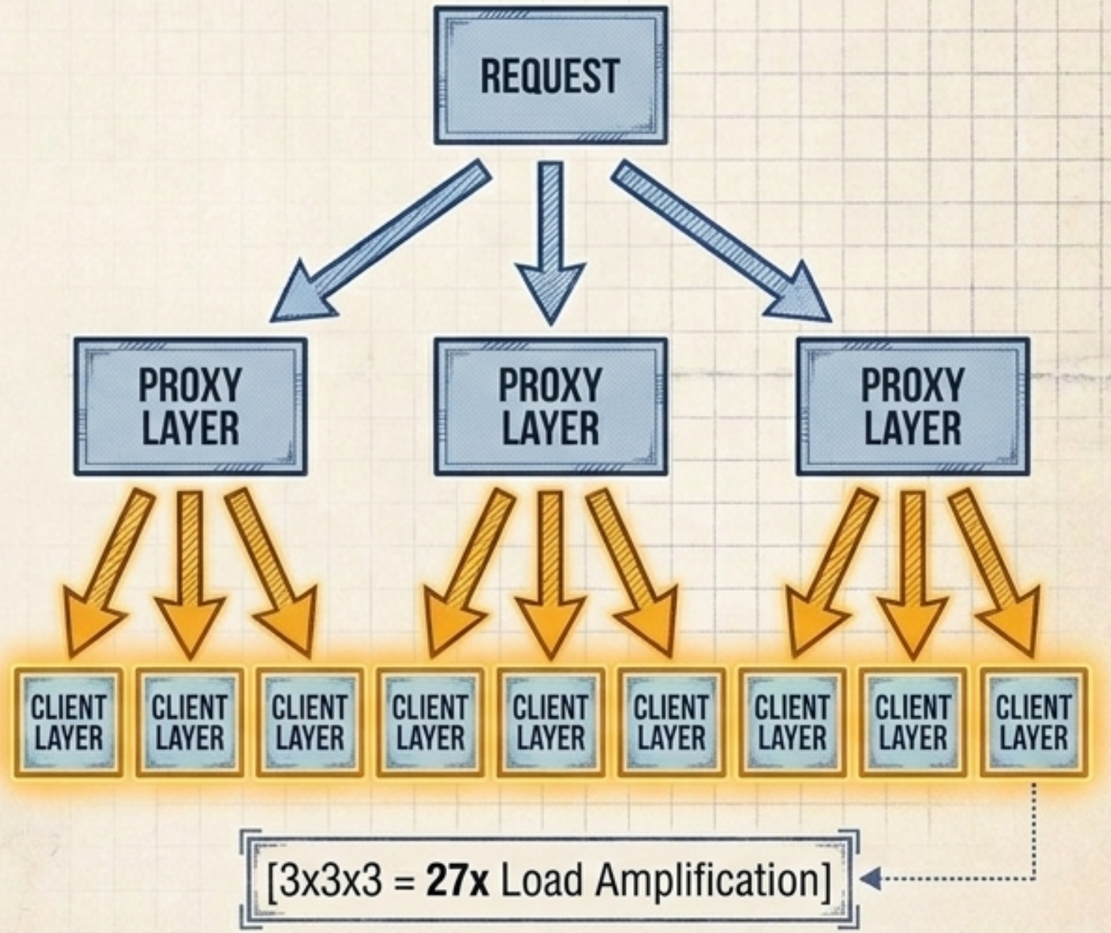
Gemini Notebook

CASCADING FAILURES & THE RETRY STORM

THE DEATH SPIRAL



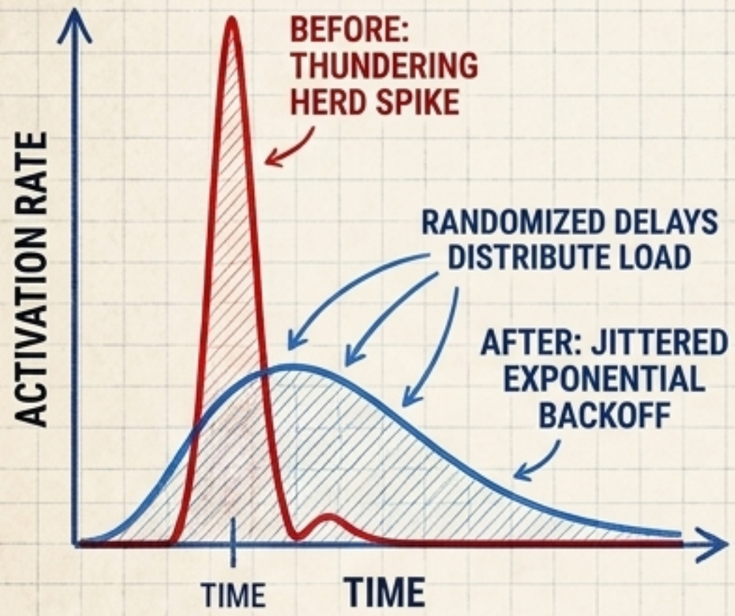
THE AMPLIFICATION THREAT



Gemini Notebook

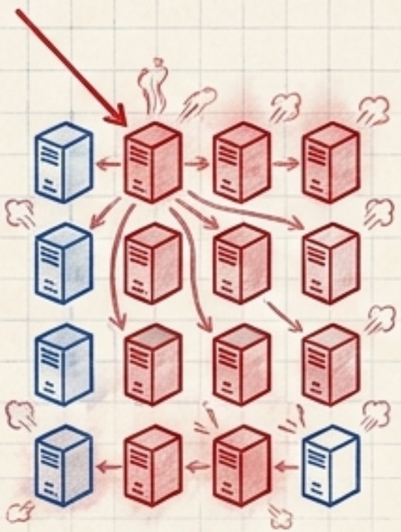
BREAKING THE CHAIN: THE RESILIENCE TOOLKIT

JITTERED BACKOFF

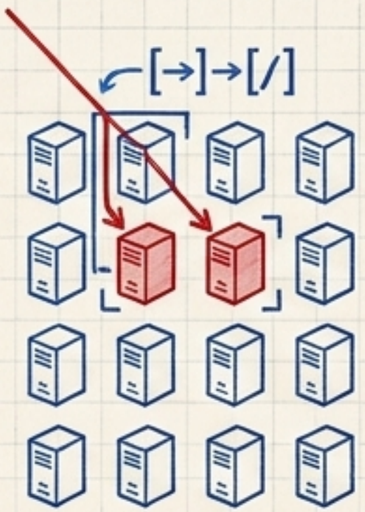


Adding randomness to exponential backoff prevents thundering herd retry spikes.

SHUFFLE SHARDING



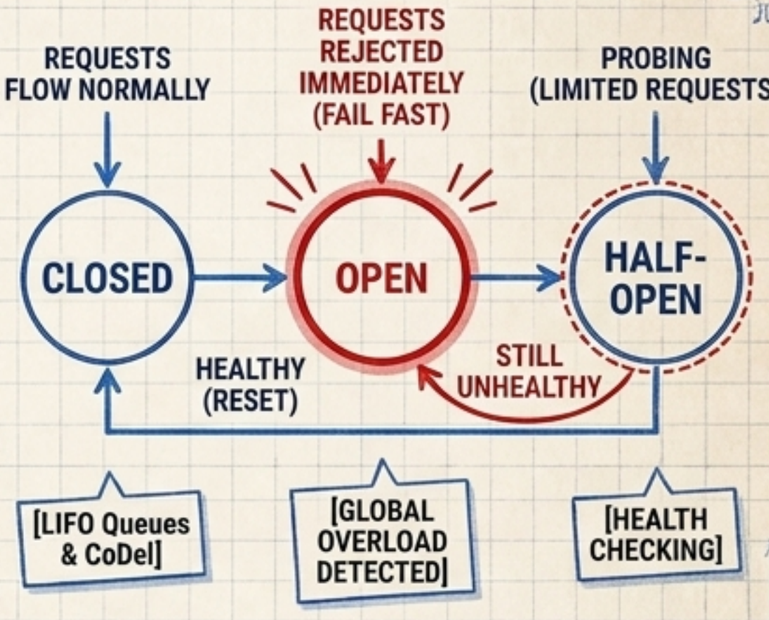
NAIVE STATE: UNCONSTRAINED



RESILIENT STATE: SHARDED ISOLATION

A ‘poison pill’ request is mathematically constrained to connect to only 2 specific nodes, leaving the rest safely blue.

LOAD SHEDDING & CIRCUIT BREAKERS



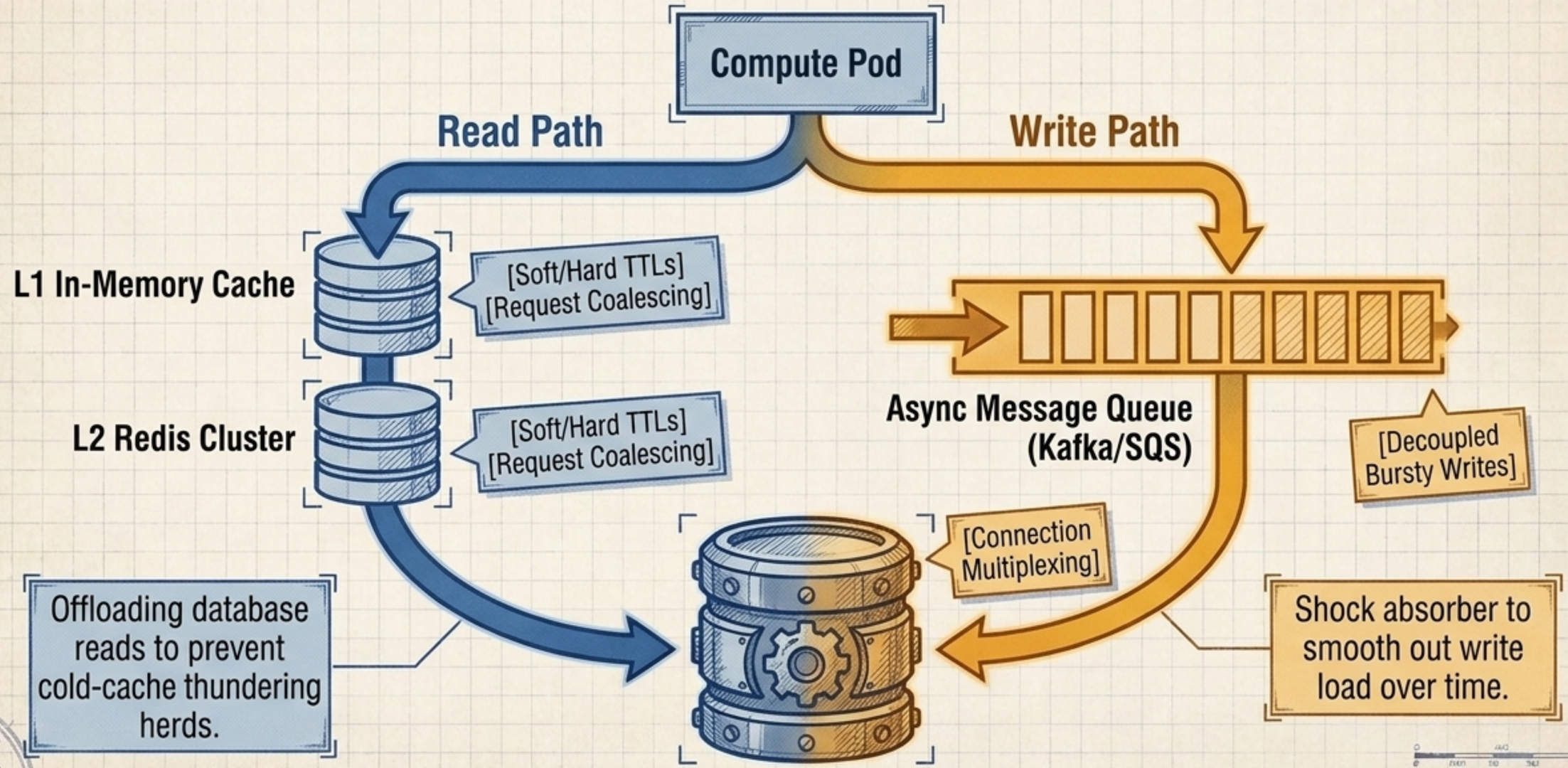
Dropping requests early via short RPC deadlines when global overload is detected

[PREVENT CASCADING FAILURE]

[PRESERVE CRITICAL SYSTEMS]

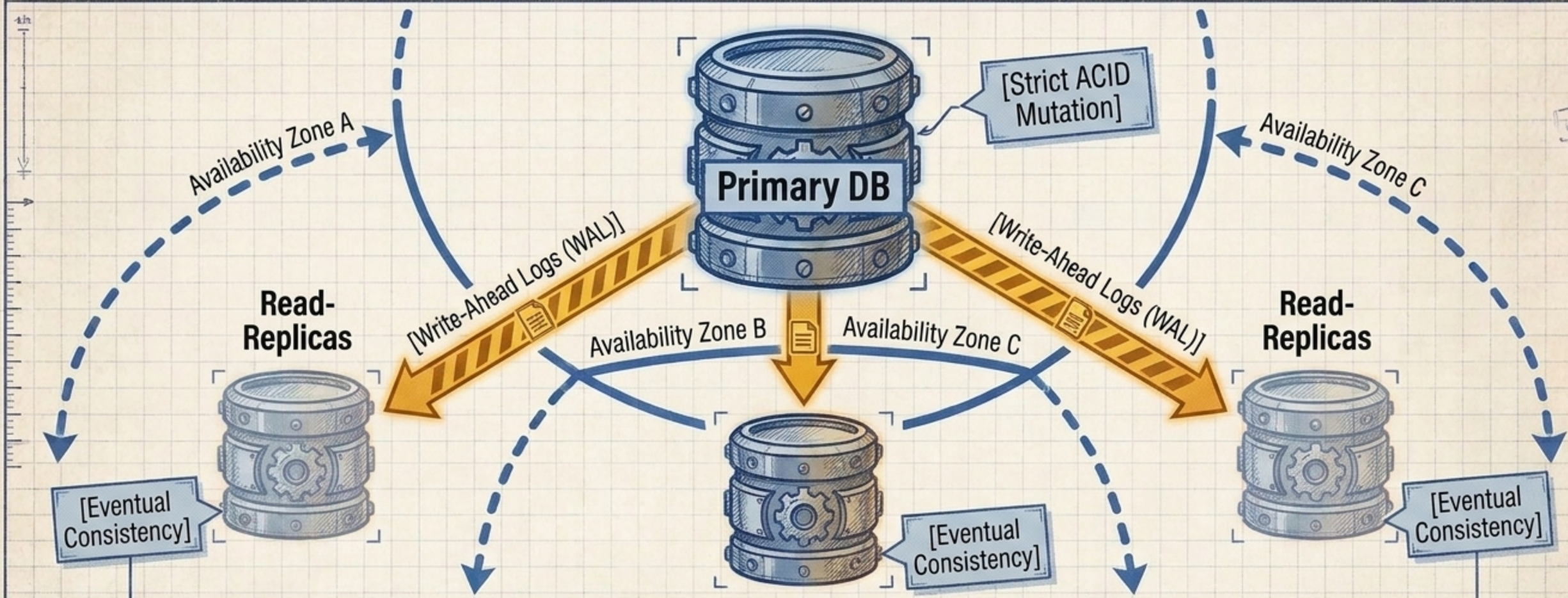
Gemini Notebook

SCALING STATE: L1/L2 CACHES & ASYNC QUEUES



Gemini Notebook

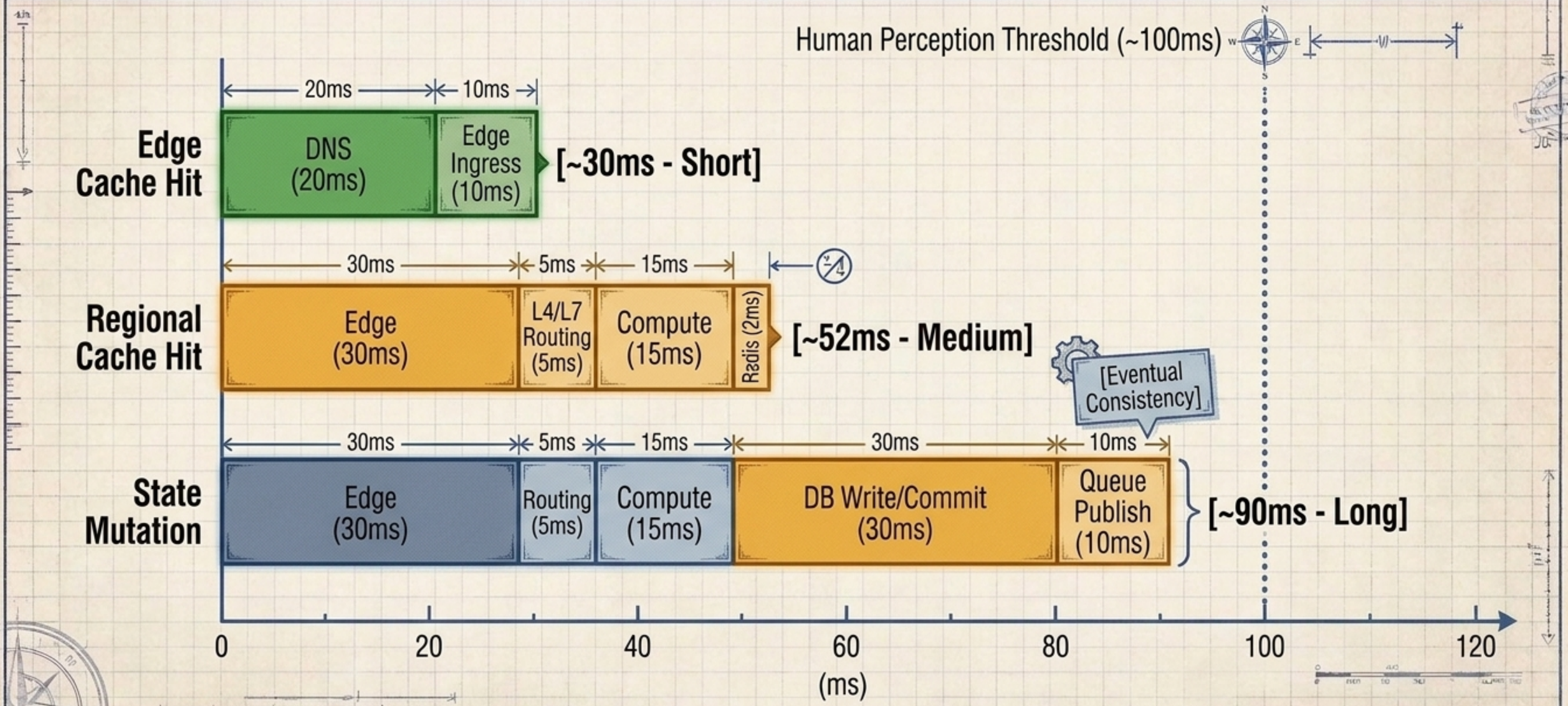
DURABLE PERSISTENCE: PRIMARY/REPLICA TOPOLOGIES



Eventual Consistency Trade-off: Routing reads to replicas risks serving stale data, but is mandatory to preserve Primary CPU threads for data integrity and write mutations.

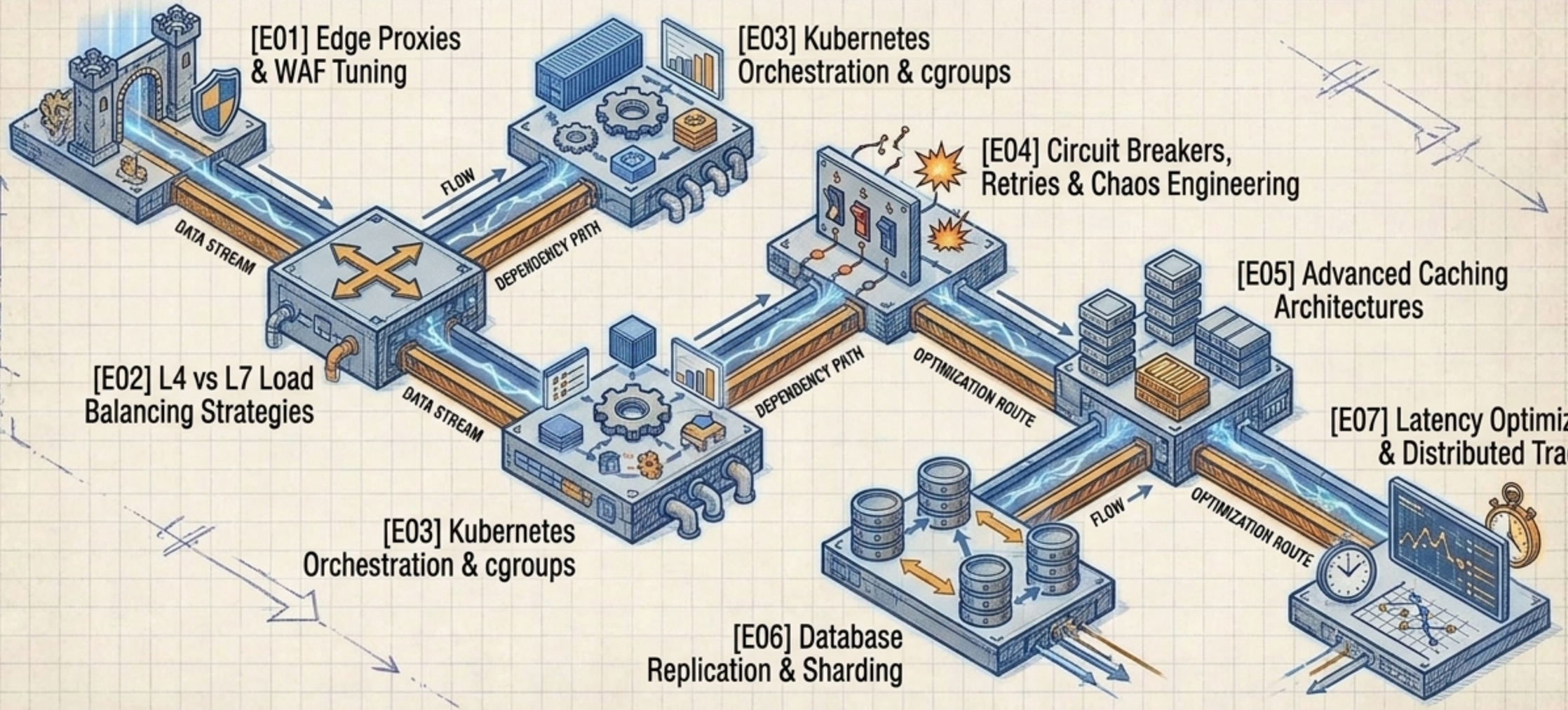
Gemini Notebook

THE MILLISECOND BUDGET: CACHE HIT VS. STATE MUTATION



Gemini Notebook

SEASON 02 DEEP-DIVE ROADMAP



The Telemetry Blueprint: Map your systems. Measure your bottlenecks. Master the stack.

Gemini Notebook