0000: 45 00 00 3c 1c 46 40 00 40 06 b1 e6 c0 a8 00 01
0010: c0 a8 00 02 00 50 00 50 00 00 00 00 00 00 00 00
0020: 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00 00
00A0: 02 01 00 00 00 00 00 00 00 00 00 00
00E0: 02 01 00 00 00 00 00 00 00 00 00 00
00F0: … … … …

Web Architecture — S02E01: DNS Internals, BGP-4, Anycast Routing & Global Server Load Balancing

From wire-level RFC mechanics to production GSLB failure modes.

0040: 02 01 00 00 00 1XP 00 00 00 00 00 00 00 00 00 00
0020: 02 01 00 00 00 1XP 00 00 00 00 00 00 00 00 00 00
0070: 02 00 00 00 00 000 00 00 00 00 00 00 00 00 00 00
0030: 02 00 00 00 00 1XF 00 00 00 00 00 00 00 00 00 00
00B0: 02 00 00 00 00 1XB 00 00 00 00 00 00 00 00 00 00
00F0: 02 01 00 00 00 000 00 00 00 00 00 00 00 00 00 00
00A0: 02 01 00 00 00 000 00 00 00 00 00 00 00 00 00 00
00B0: 02 01 00 00 00 1Bc 00 00 00 00 00 00 00 00 00 00
00C0: 02 01 00 00 00 001 00 00 00 00 00 00 00 00 00 00
00F0: 02 01 00 00 00 000 00 00 00 00 00 00 00 00 00 00
00F0: … … … … … … … … … … … … … … … …



NODE EDGE
DATA FLOW
HIGHLIGHTED PATH

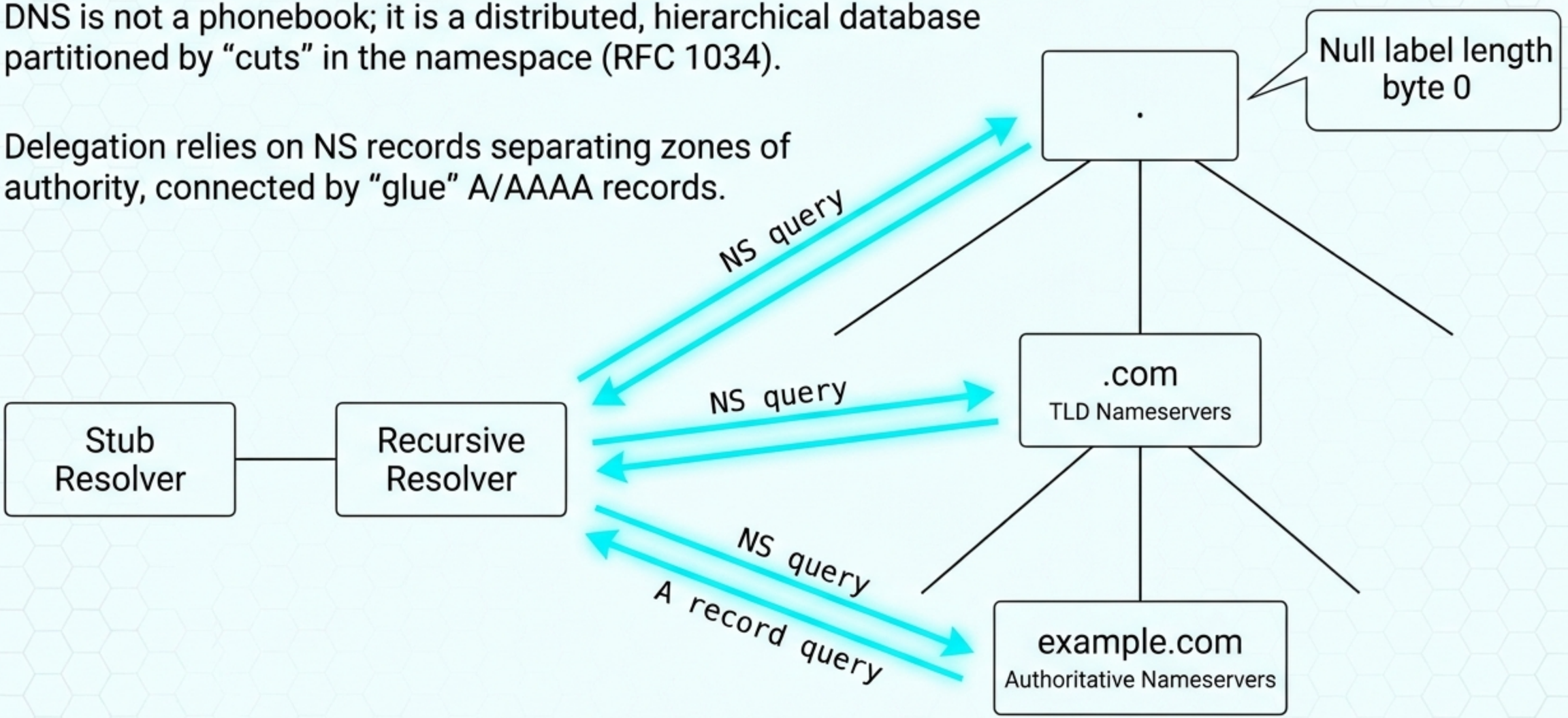
COMPONENT: WEB_ARCH_S02E01_DOC
DATE: 2024.10.27
DRAWN BY: SYSTEM_ARCHITECT_AI

Gemini Notebook

The Distributed Hierarchical Datastore

DNS is not a phonebook; it is a distributed, hierarchical database partitioned by “cuts” in the namespace (RFC 1034).

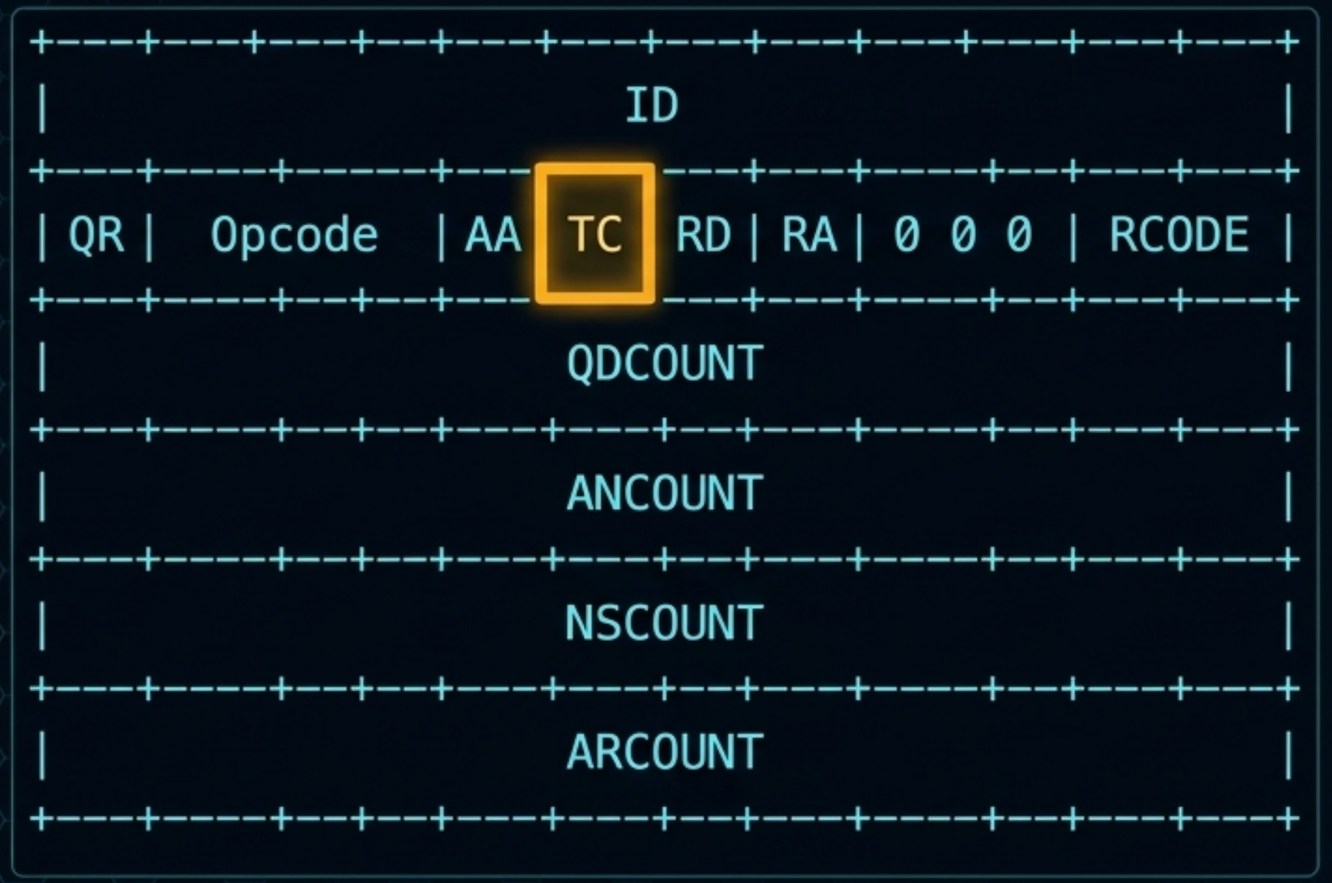
Delegation relies on NS records separating zones of authority, connected by “glue” A/AAAA records.

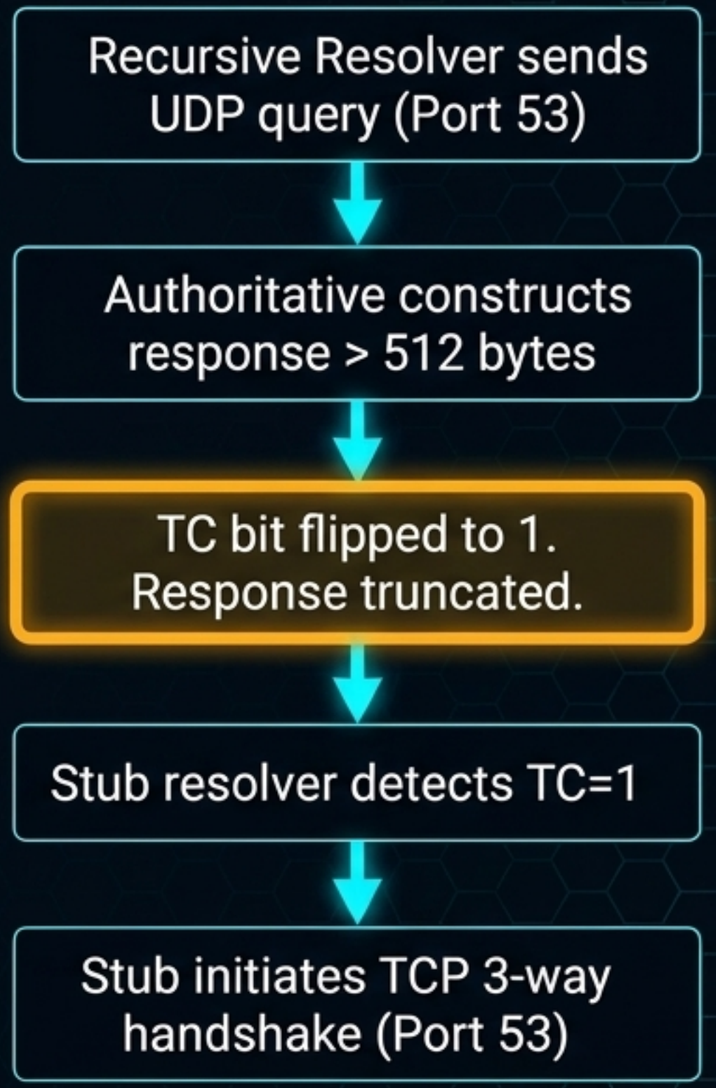


Gemini Notebook

RFC 1035 Wire Transport & The 512-Byte Limit

RFC 1035 sets a strict 512-byte limit on UDP DNS responses. Exceeding this triggers the TC bit, forcing an expensive TCP fallback to retrieve the full payload.



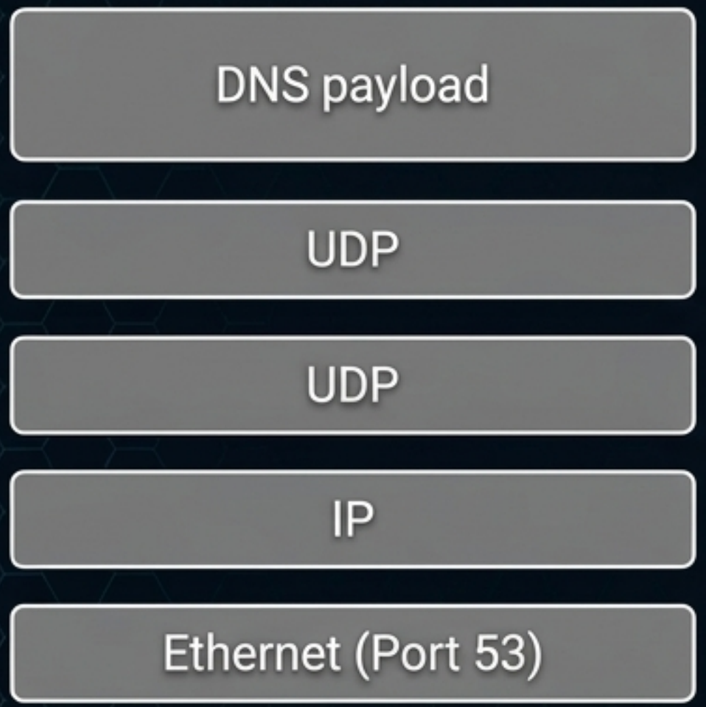


Gemini Notebook

Modern Privacy & Transport: DoT and DoH

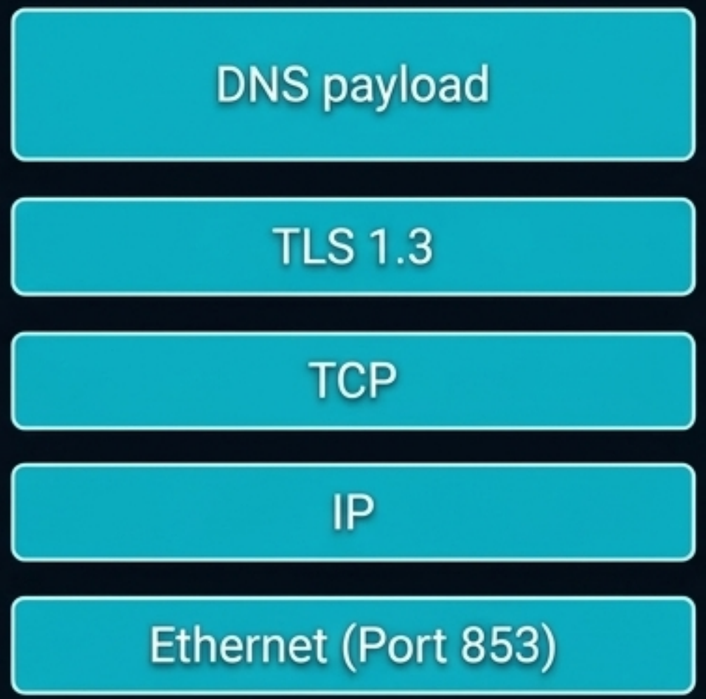
Classic DNS

Plaintext transport, zero privacy.



DNS-over-TLS (DoT)

Prevents on-path passive surveillance but uses a dedicated port (853), making it trivially blockable by firewalls.



DNS-over-HTTPS (DoH)

Encapsulates wire-format DNS messages in base64url via HTTP GET/POST, blending indistinguishably with standard web traffic on Port 443. Limit expanded to 65535 bytes.



Gemini Notebook

The DoH Centralization Trade-off

DoH solves the last-mile surveillance problem but alters the internet's topology. It routes global DNS traffic away fram from edge ISPs into a handful of centralized megacache honeypots, creating massive single points of failure.

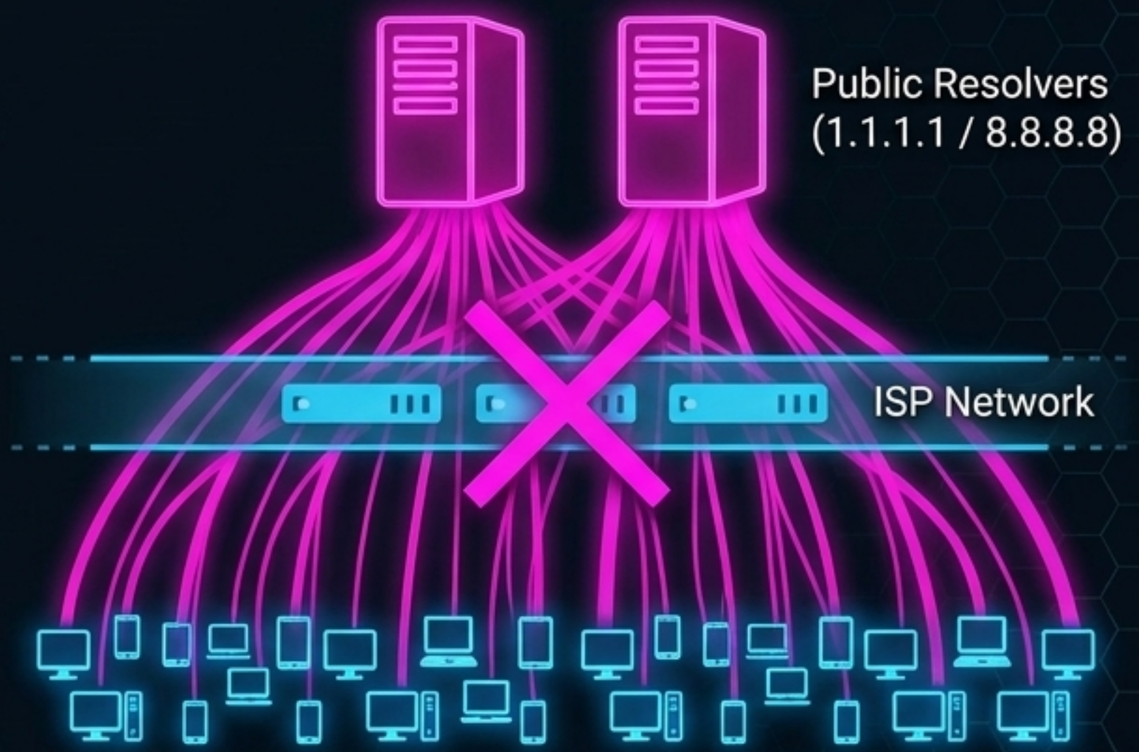
Traditional ISP Resolution



Client Endpoint

Distributed, decentralized, vulnerable to local ISP logging.

DoH Centralization



Client Endpoint

Single points of failure, massive traffic aggregation.

Gemini Notebook

Record Semantics & Negative Caching

```
example.com.   IN   SOA ns1.example.com. admin.example.com. (
                      2023102401   ; SERIAL
                      7200         ; REFRESH
                      3600         ; RETRY
                      1209600      ; EXPIRE
                      300 )        ; MINIMUM (Negative Cache TTL)
```

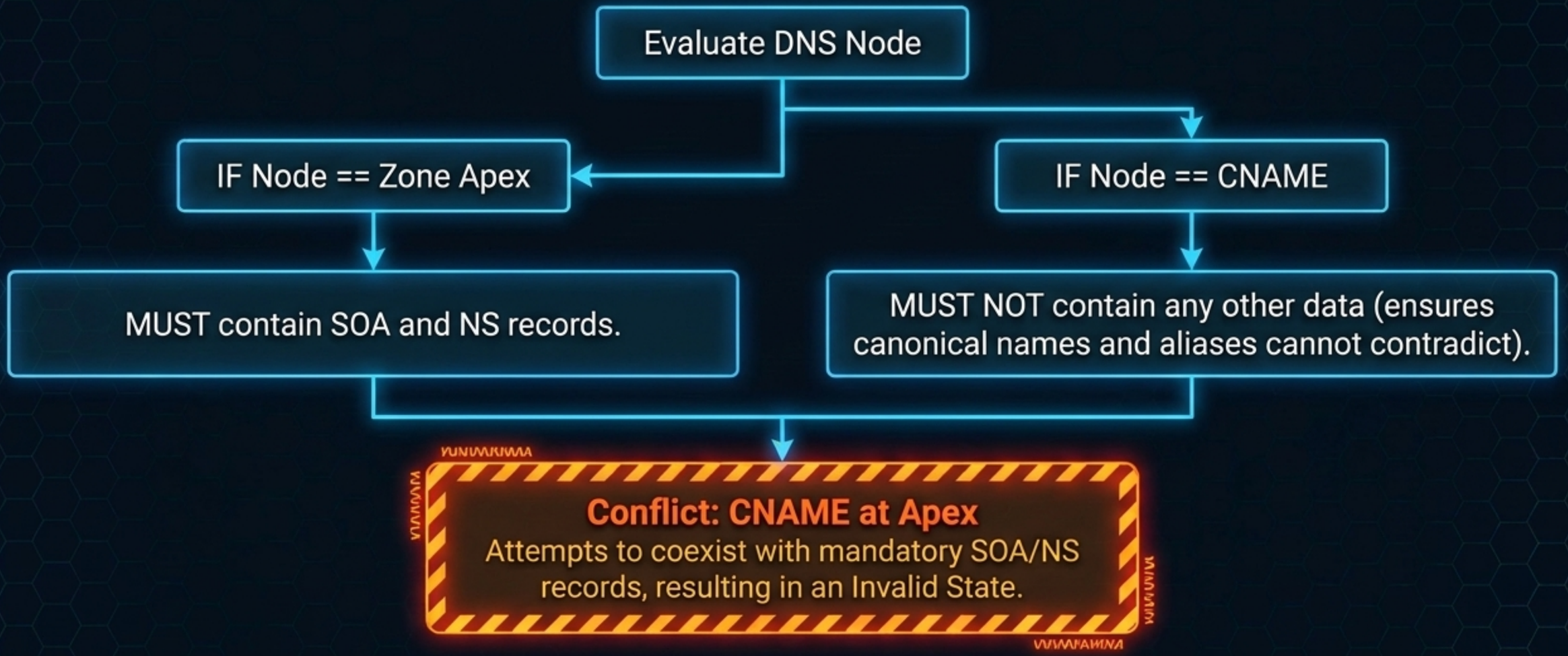
Governs Negative Caching (RFC 2308)

If a resolver queries a non-existent subdomain, the authoritative server returns NXDOMAIN with this TTL. The resolver caches the absence of the record, preventing query floods.

Gemini Notebook

The Zone Apex CNAME Constraint (RFC 1034)

RFC 1034 Rule 3.6.2: A CNAME record identifies its owner name as an alias. If a CNAME RR is present at a node, no other data should be present. Ergo, CNAME at the Apex is an illegal state.

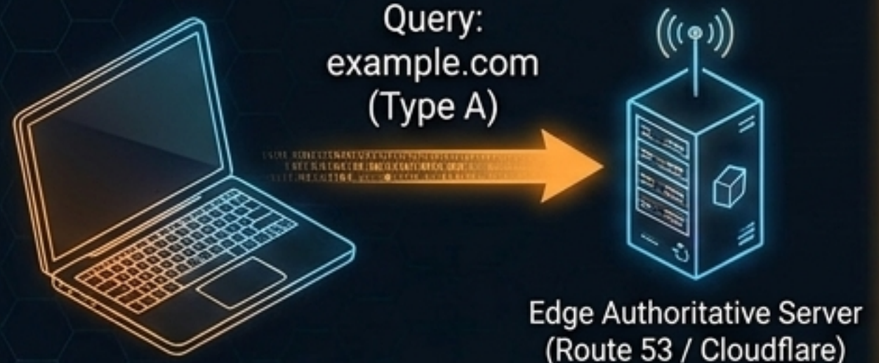


Gemini Notebook

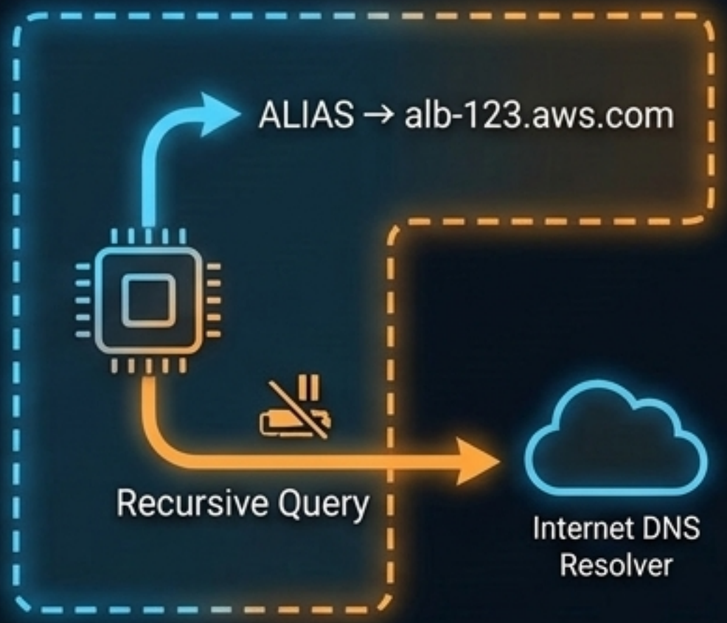
Overcoming the Apex Constraint: Record Flattening

To beat the RFC 1034 constraint, modern DNS providers implement "Flattening". The authoritative edge internally resolves the alias and synthesizes a dynamic A/AAAA response, entirely hiding the CNAME from the querying client.

The Client Query



Internal Resolution



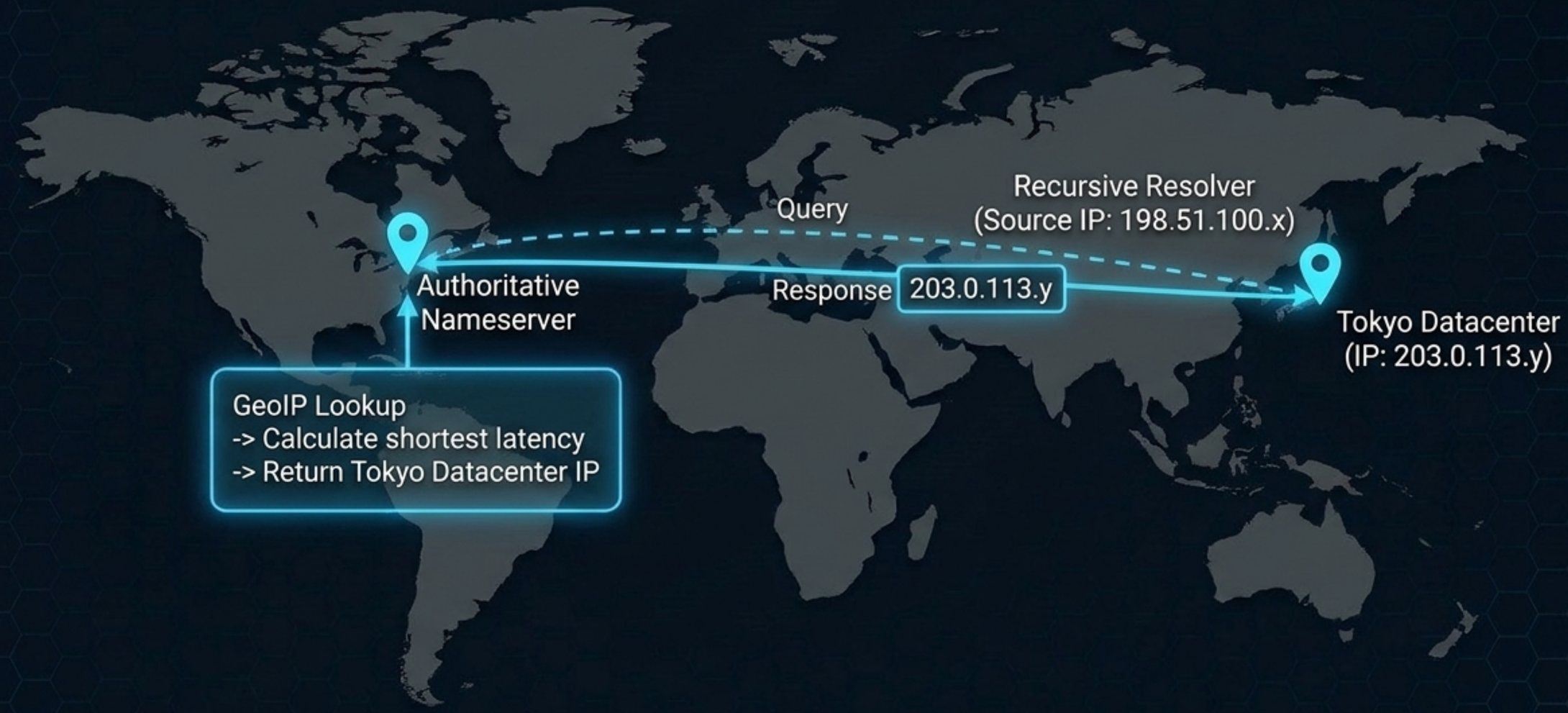
Synthesis & Response



Gemini Notebook

Unicast GeoDNS Steering Mechanics

Unicast GeoDNS Steering evaluates the incoming resolver's source IP, calculates optimal latency, and dynamically alters the A/AAAA response to steer traffic to the closest regional PoP.



Gemini Notebook

The Centralized Resolver Blindspot

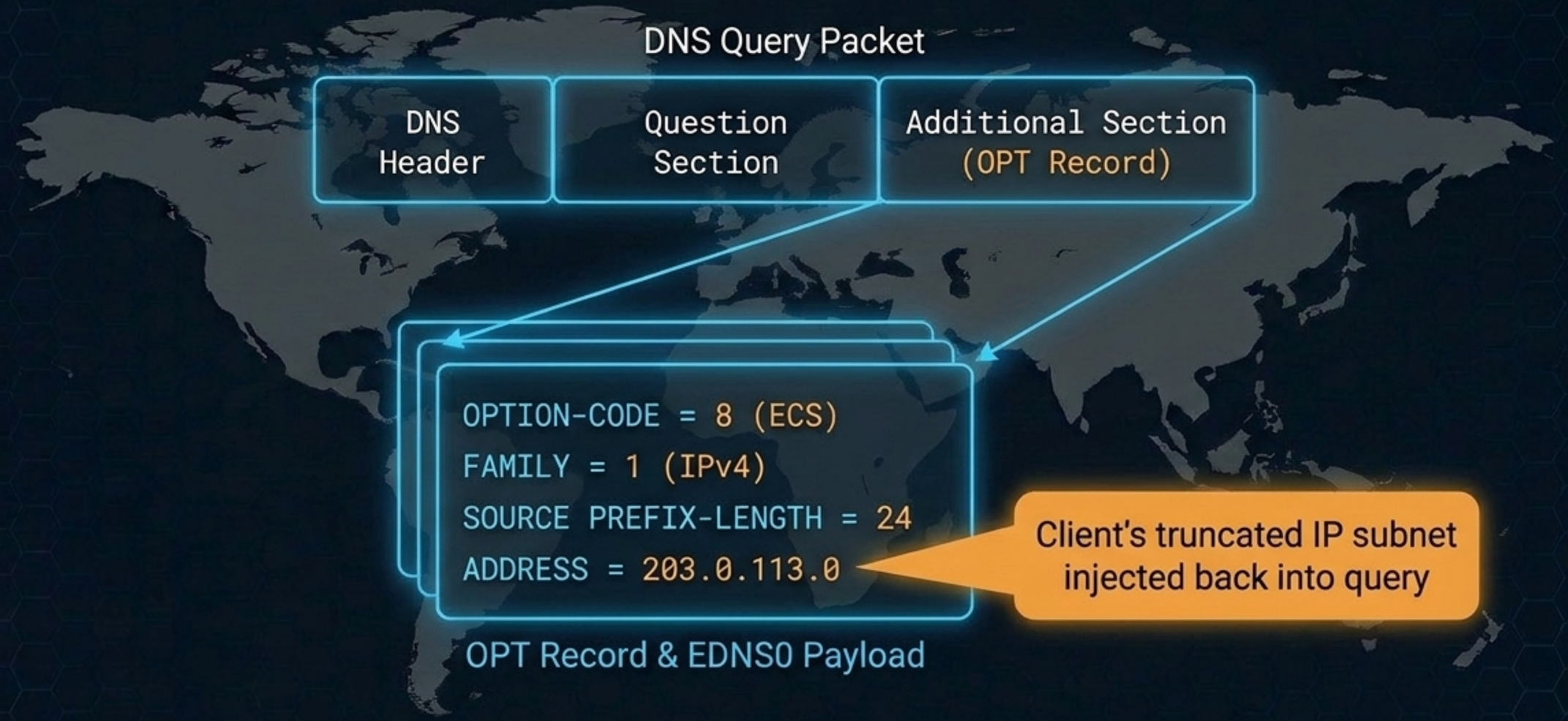
Centralized public resolvers obscure the true client origin. The authoritative server optimizes for the resolver's location, inadvertently steering edge users to suboptimal distant datacenters.



Gemini Notebook

EDNS Client Subnet (ECS) to the Rescue

EDNS Client Subnet (ECS) allows the recursive resolver to pass a truncated portion of the actual client's IP inside the query's EDNS0 payload. The authoritative server can now steer based on true user geography.

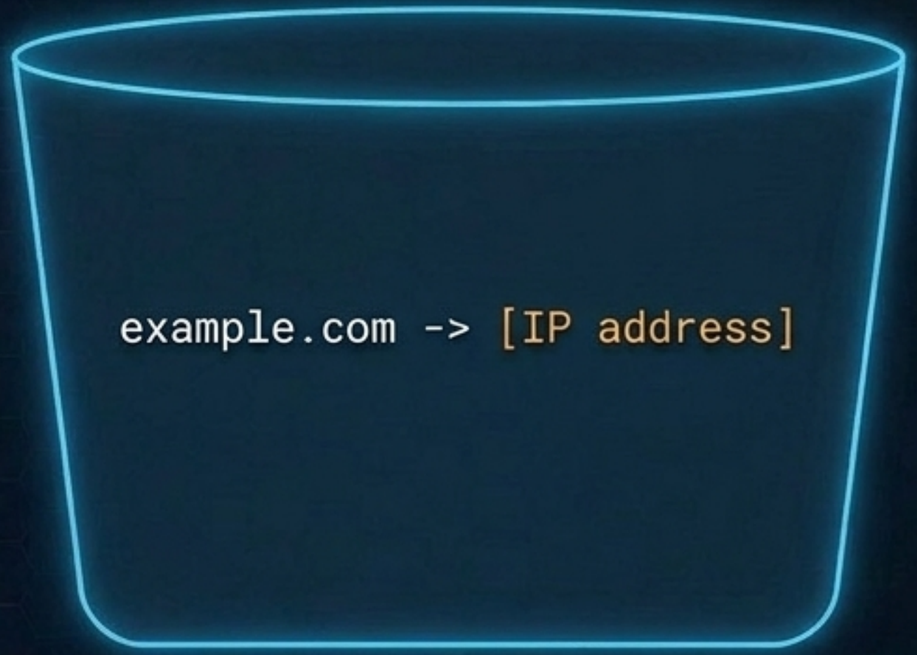


Gemini Notebook

ECS Trade-offs: Privacy vs. Cache Fragmentation

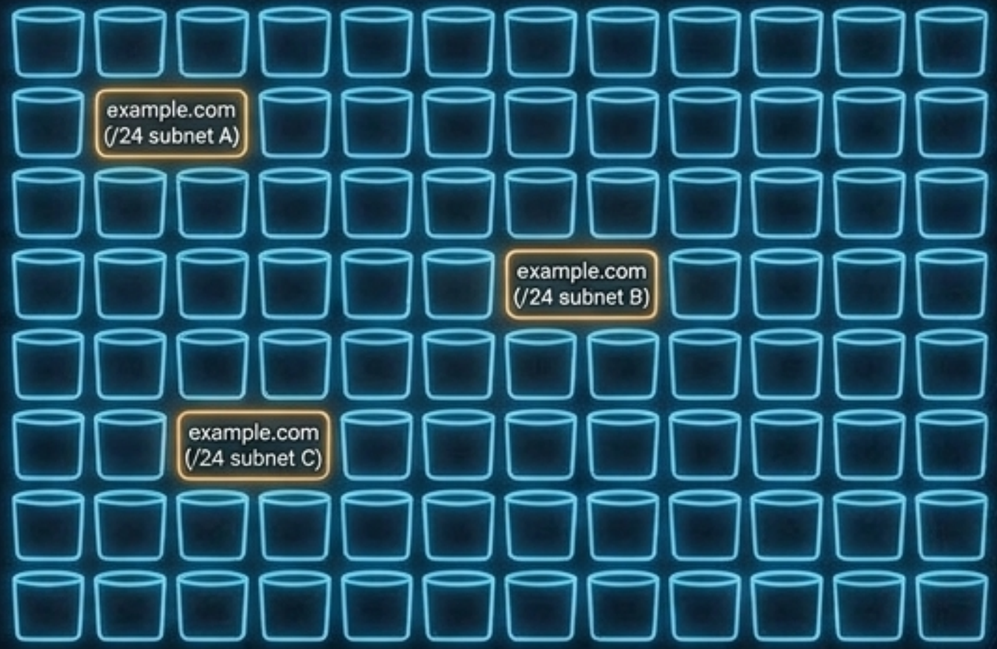
ECS destroys cache efficiency. Instead of caching one response for a domain, the recursive resolver must store a separate entry for every unique /24 subnet that requests it. This cache fragmentation increases memory overhead and tanks cache hit rates.

Standard DNS Caching (No ECS)



High cache hit rate,
low memory overhead.

ECS Caching

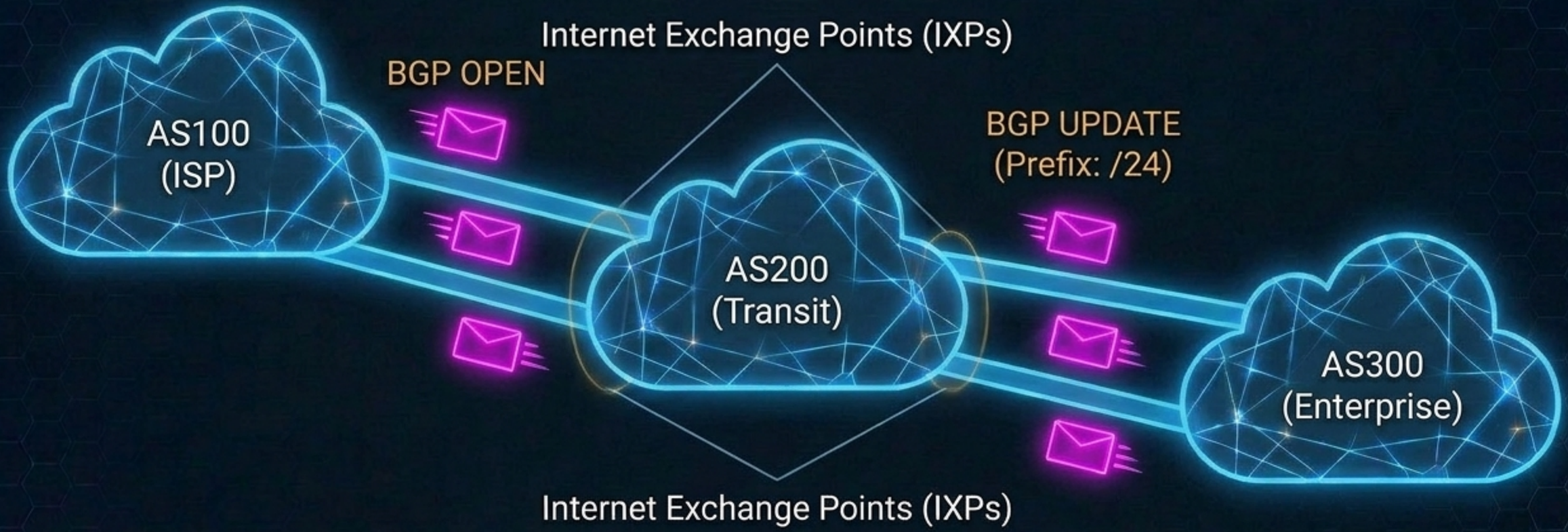


Severe fragmentation.
Poor hit rates.

Gemini Notebook

BGP-4 & Autonomous Systems (ASNs)

The internet is a graph of independent Autonomous Systems (ASNs). BGP-4 routes traffic strictly on AS_PATH length (fewest network hops)—completely ignoring actual bandwidth or geographic distance.



Gemini Notebook

The Magic of Anycast Routing

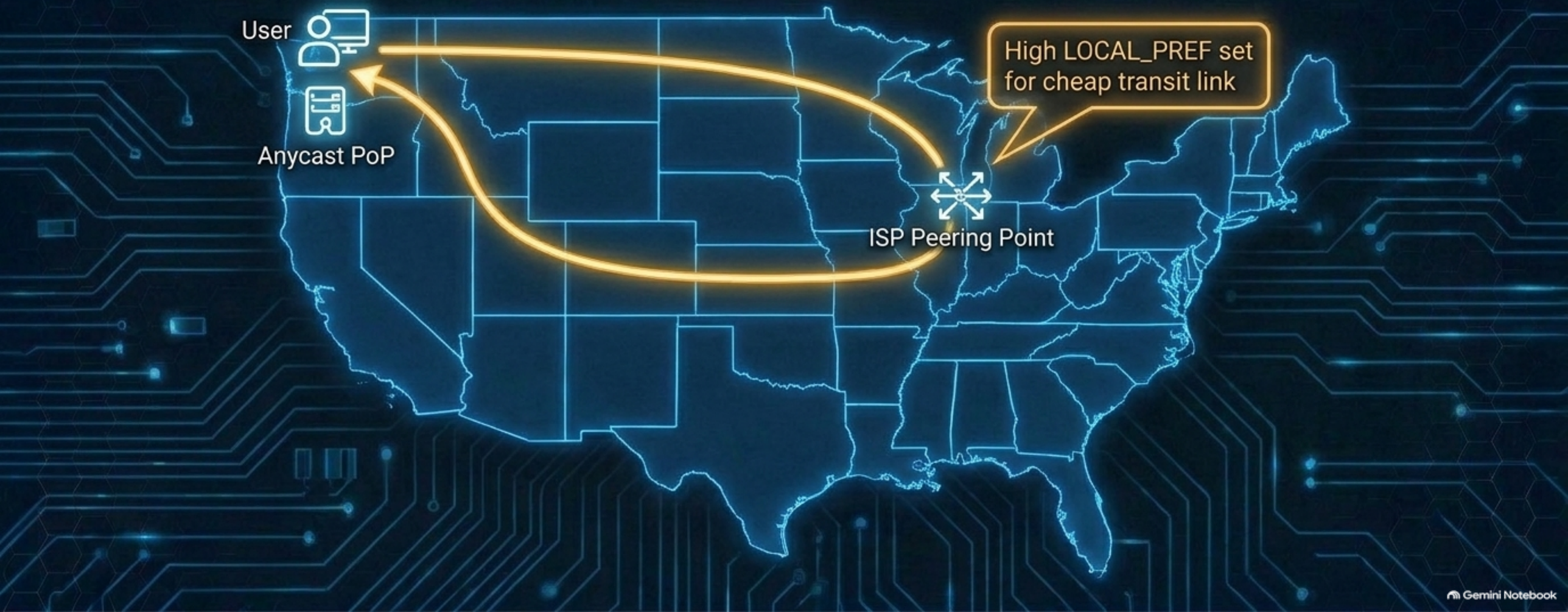
In Anycast, multiple datacenters advertise identical IP prefixes globally. The internet's BGP routing tables automatically converge to send client packets to the topologically closest healthy location.



Gemini Notebook

The ‘BGP Hairpinning’ Anomaly

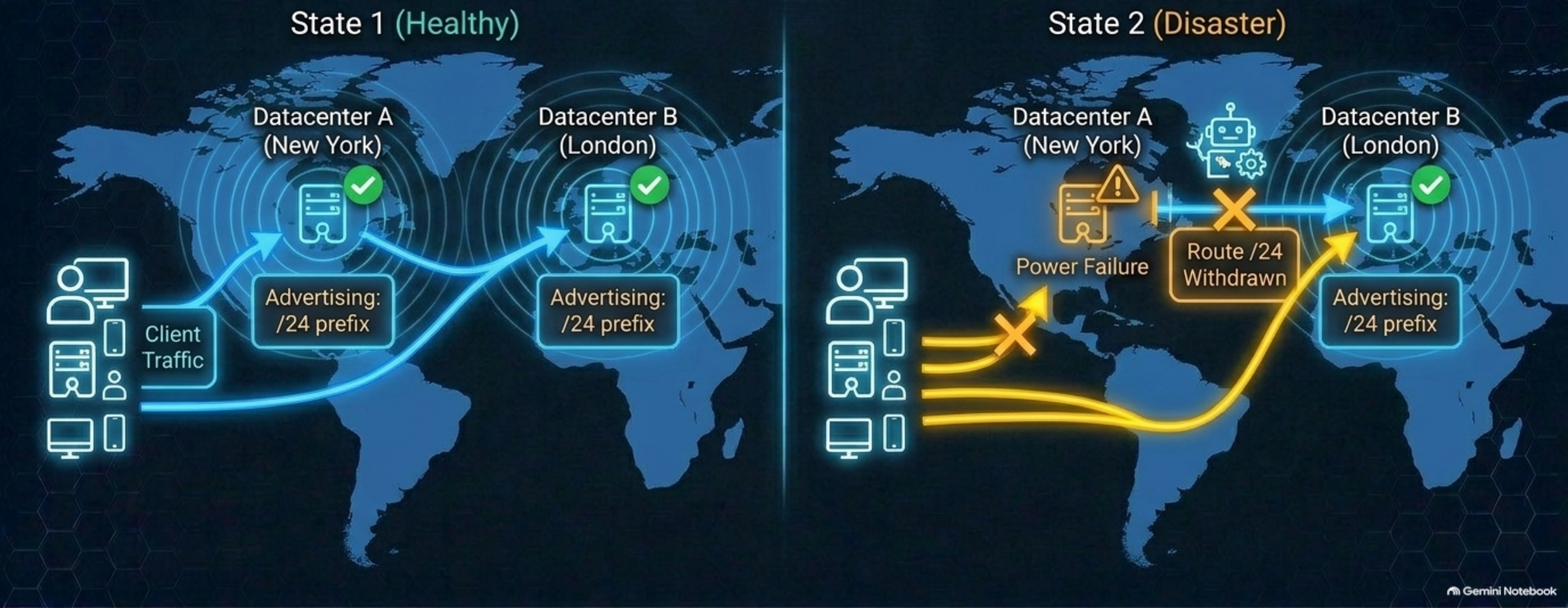
BGP logic trumps geographic reality. ISP peering policies (LOCAL_PREF, MED) frequently override topological proximity, forcing packets to ‘hairpin’ across continents to reach a datacenter that is physically next door.



Gemini Notebook

Outage Mitigation via BGP PoP Withdrawal

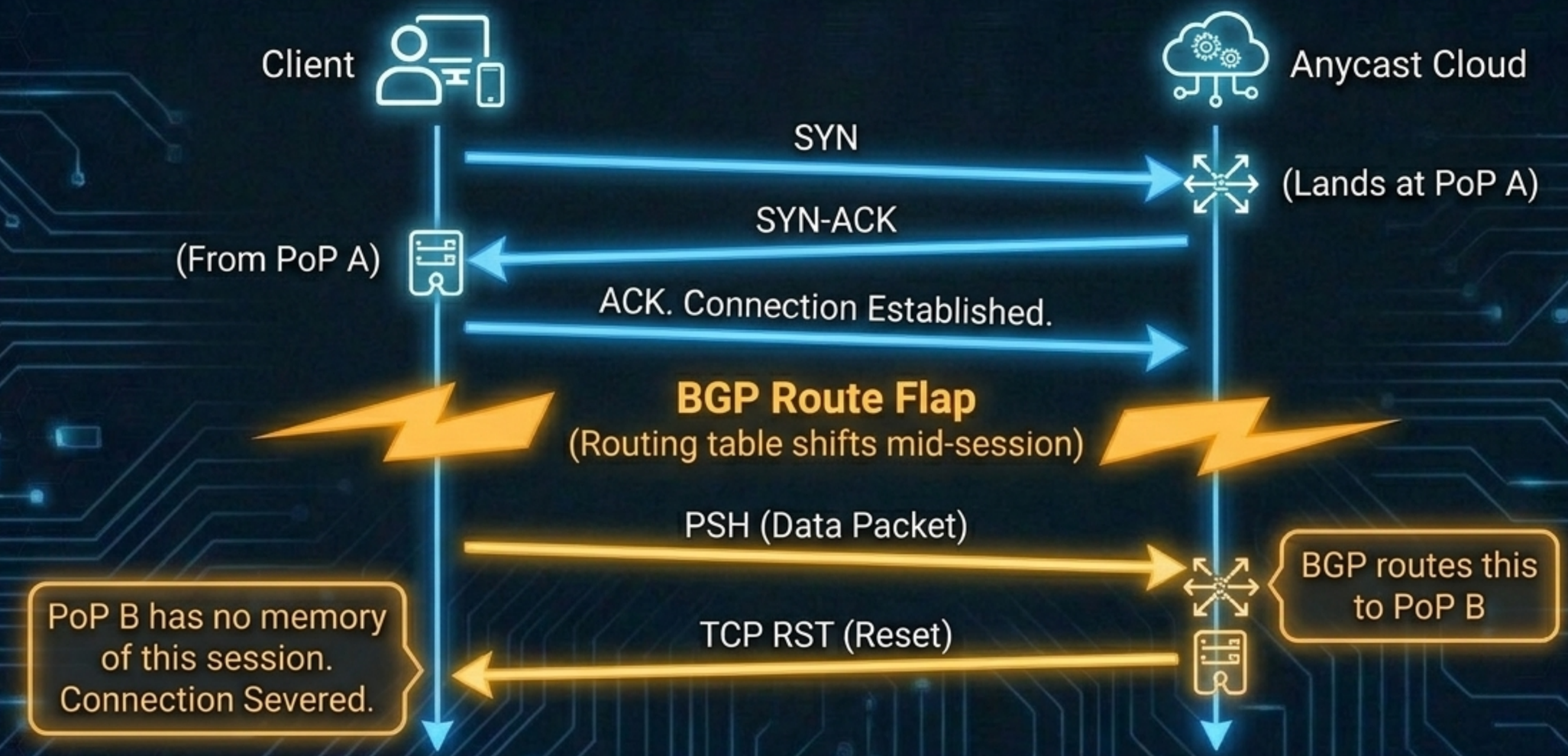
Outage mitigation at Layer 3: When a PoP fails its internal health checks, it simply withdraws its BGP route. The internet automatically reconverges around the failure, shifting packets to the next closest PoP.



Gemini Notebook

The Threat of Route Churn to Stateful Sessions

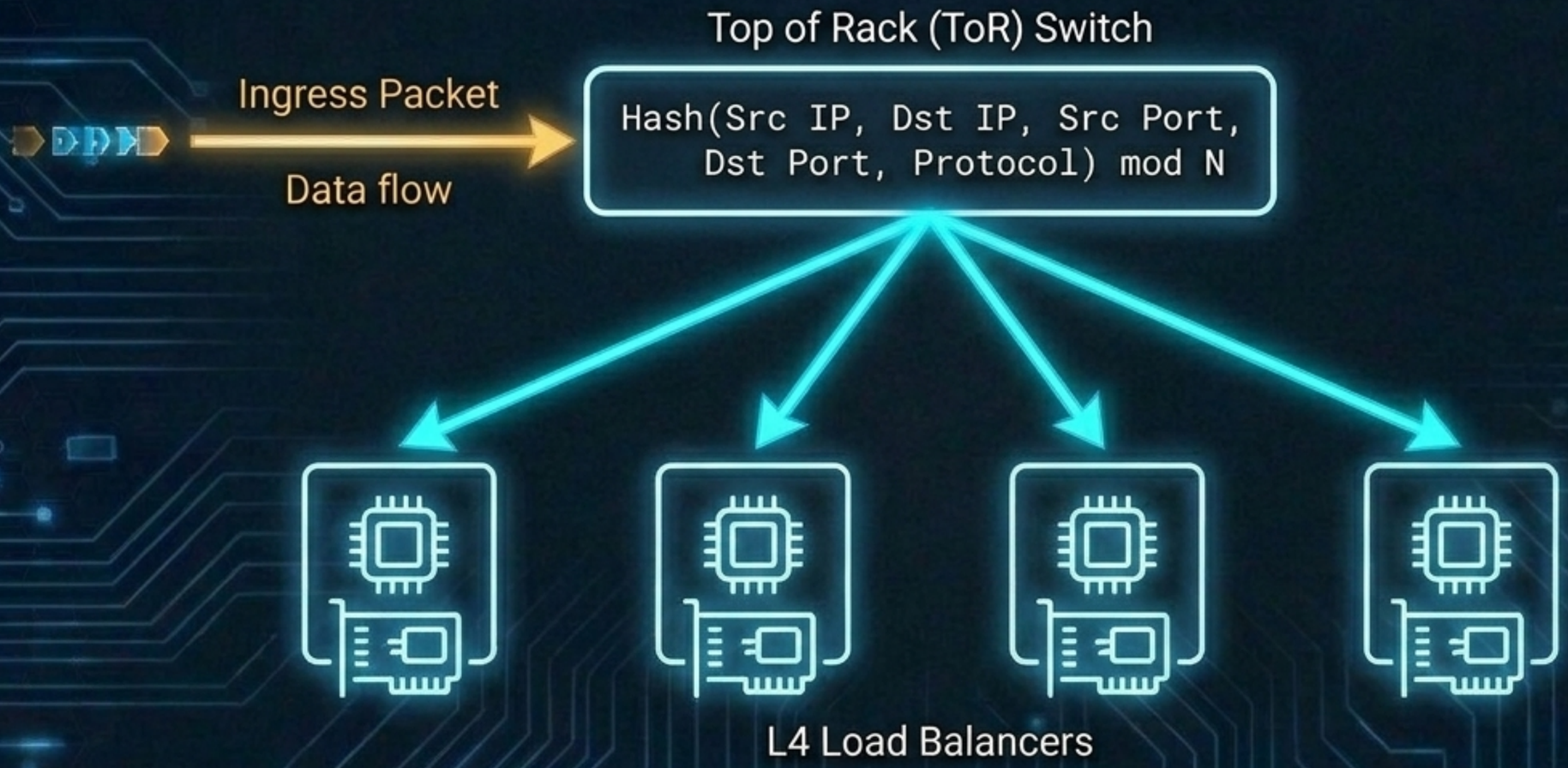
BGP route churn is lethal to TCP/QUIC. If the routing path shifts mid-flight, subsequent packets in a stateful session will land at a different datacenter, resulting in immediate TCP resets and dropped connections.



Gemini Notebook

PoP-Internal ECMP (Equal-Cost Multi-Path)

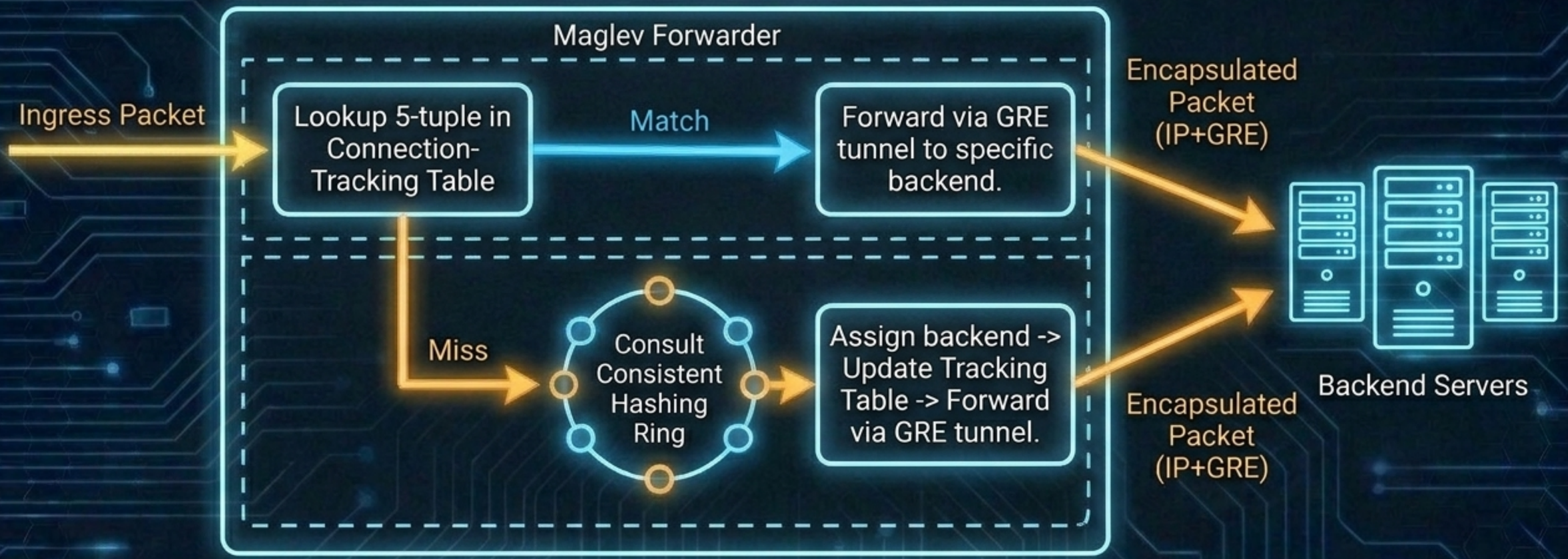
PoP-Internal ECMP routing utilizes a 5-tuple hash to distribute incoming packets across a fleet of internal load balancers. However, if backend count 'N' changes due to failure, the hash re-maps, threatening all active connections.



Gemini Notebook

Connection-Tracking & Consistent Hashing (Maglev)

To survive backend churn, forwarders use Connection-Tracking combined with Consistent Hashing. Packets are encapsulated (IP+GRE) and pinned to the correct backend, ensuring stateful sessions survive even when individual servers fail.



Gemini Notebook

Control Plane vs. Data Plane Separation

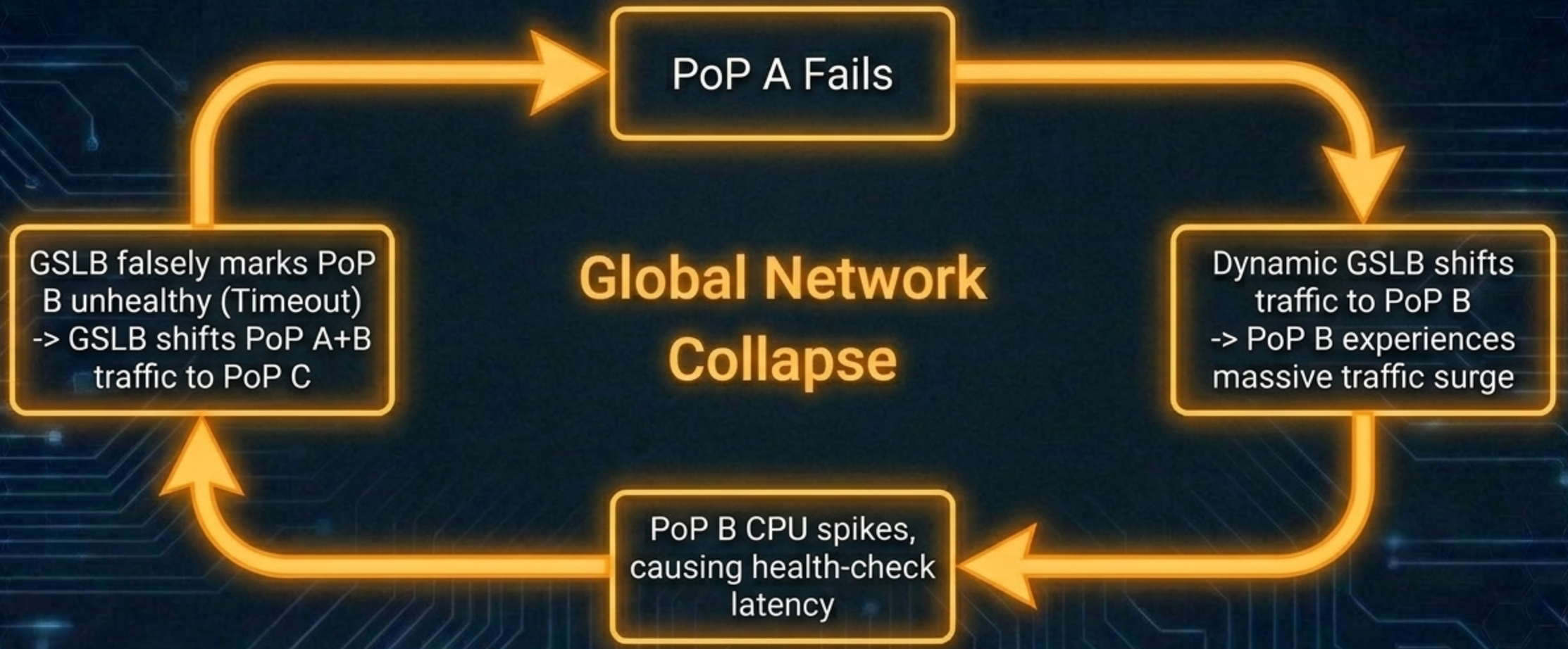
Reliability requires physical and logical isolation. The Control Plane evaluates health and configures routes (slow, complex). The Data Plane blindly forwards packets based on the current configuration (stateless, microsecond latency).



Gemini Notebook

The Danger of Dynamic Reconfiguration Cascades

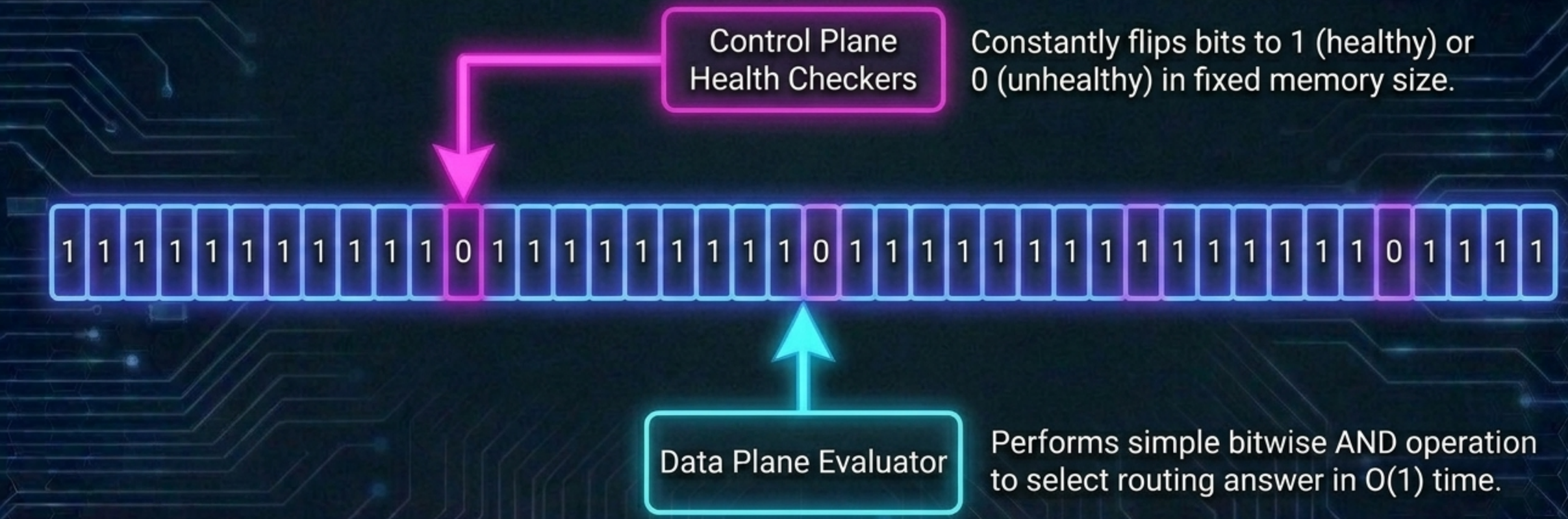
Dynamic failover based on live system load is a trap. It creates cascading failures where traffic shifts trigger artificial health check timeouts, methodically taking down healthy datacenters one by one.



Gemini Notebook

The Constant-Work Pattern (AWS Route 53 Design)

Engineering for extreme reliability requires Constant-Work algorithms. The system performs the exact same bitwise operations and consumes the exact same CPU/memory regardless of whether 0% or 100% of endpoints are failing. 100% deterministic latency; zero dynamic scaling during a crisis.



Gemini Notebook

Production Failure: The “Rogue Caching” Problem

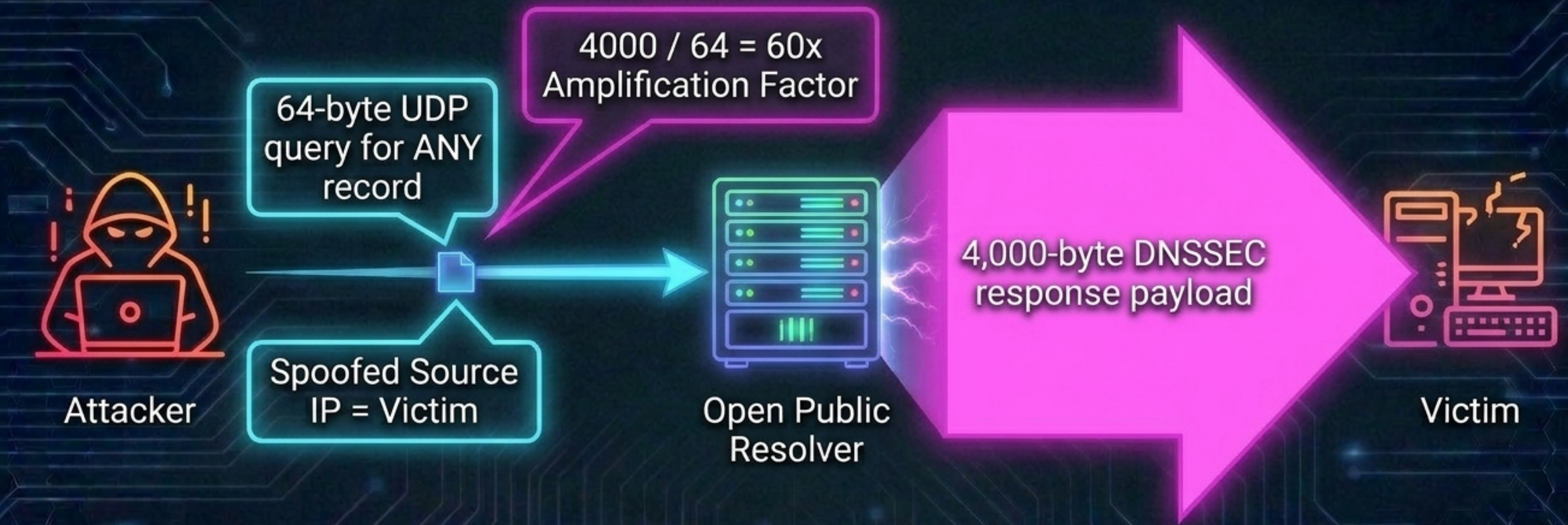
DNS TTLs are treated as suggestions by intermediate resolvers and are routinely ignored by application connection pools (e.g., default JVMs). DNS-based failover is powerless to rescue legacy clients holding stale records.



Gemini Notebook

Production Failure: DNS Amplification DDoS

Because UDP is stateless, attackers can spoof a victim's source IP. By requesting massive DNSSEC records, they trick open resolvers into bombarding the victim with 60x more bandwidth than the attacker generated.



Gemini Notebook

Production Failure: BGP Route Hijacking

Anycast prefixes can be stolen by misconfigured or malicious ASNs. Mitigation requires strict RPKI (Resource Public Key Infrastructure) ROAs to cryptographically validate route origins before propagation.



Gemini Notebook

The GSLB Decision Matrix

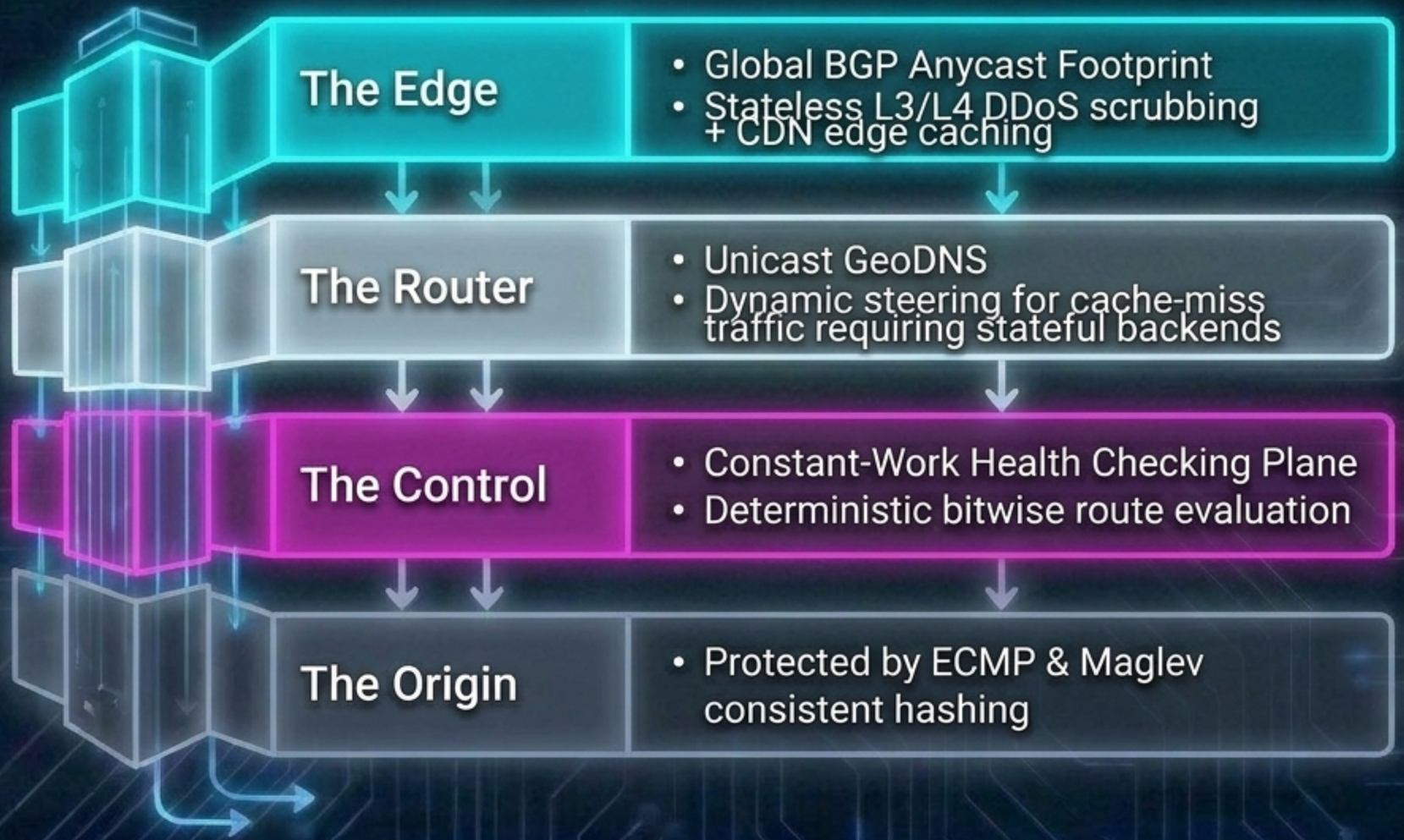
No single GSLB solution handles all constraints. Anycast wins on convergence speed and DDoS scrubbing, but Unicast GeoDNS is mandatory for steering long-lived, stateful TCP connections.

	Anycast Edge	Unicast GeoDNS (Latency)	Weighted DNS
Convergence Time	Milliseconds	Minutes	Minutes
Stateful Protocol Compatibility	Poor / Vulnerable to Churn	Excellent	Excellent
DNS TTL Reliance	None	High	High
Cache Fragmentation	None	High (ECS driven)	Severe

Gemini Notebook

Executive Synthesis: The GSLB Stack

The Modern GSLB Stack: BGP Anycast absorbs the internet's chaos at the edge. GeoDNS steers dynamic traffic to precise regional datacenters. Maglev preserves state. **Constant-work** control planes ensure deterministic survival during catastrophic failures.



Gemini Notebook