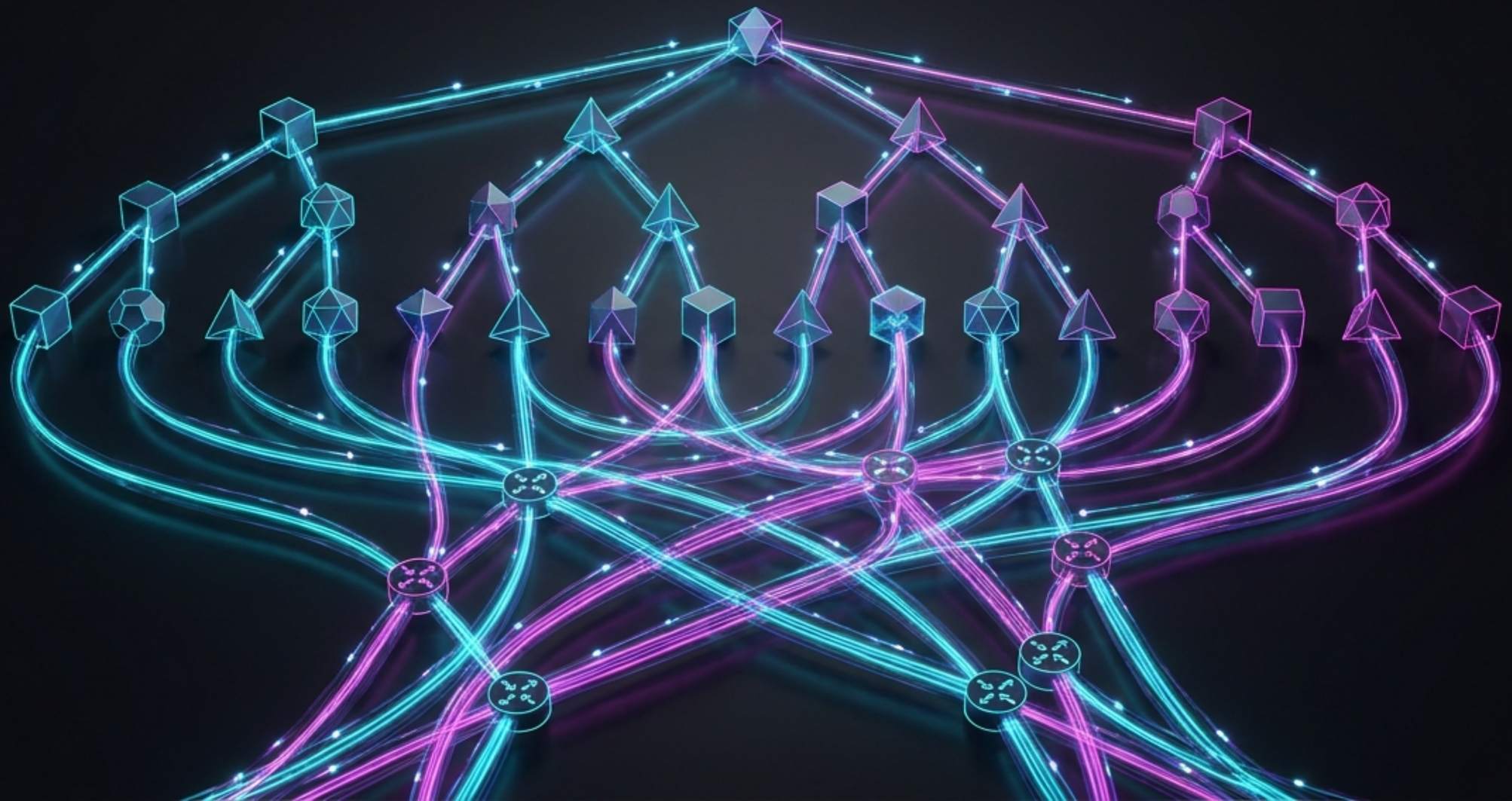
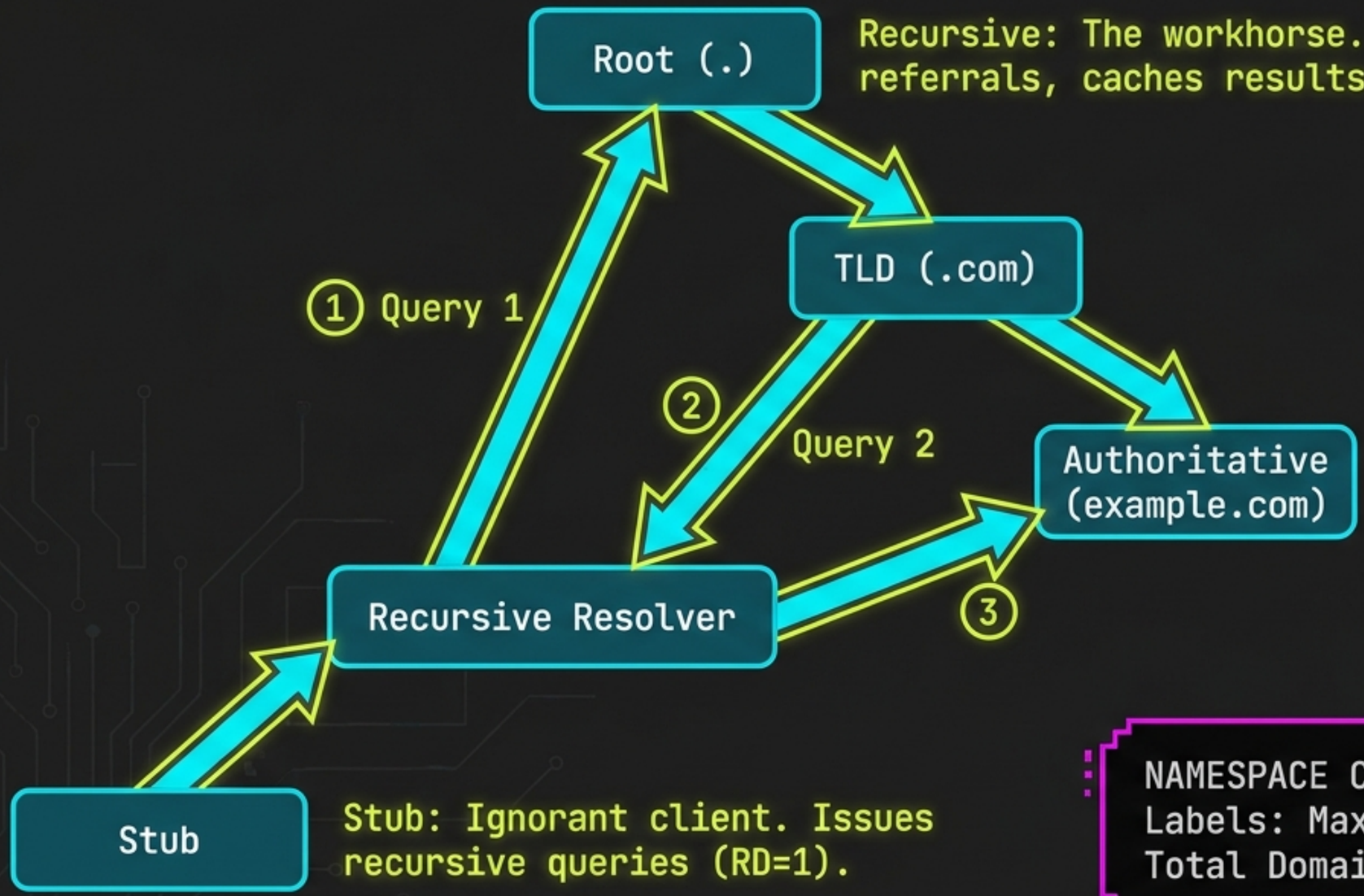
# Web Architecture — S02E01: DNS Internals, BGP-4, Anycast Routing & Global Server Load Balancing

STATUS: 200 OK | PROTOCOL: RFC-1034/1035/4271/8484



Gemini Notebook

# THE DISTRIBUTED NAMESPACE TREE (RFC 1034)




Recursive: The workhorse. Follows referrals, caches results.

Authoritative: Source of truth. Operates in discrete zones cut along the namespace.

Stub: Ignorant client. Issues recursive queries (RD=1).

NAMESPACE CONSTRAINTS (RFC 1034/1035):
Labels: Max 63 octets
Total Domain Name: Max 255 octets

Gemini Notebook

# On the Wire: RFC 1035 & UDP Transport

Legacy UDP DNS is capped at 512 bytes (excluding IP/UDP headers). When an authoritative response exceeds this boundary, the server sets the TC (Truncation) bit to 1.

## DNS Header Format (16-Bit)

hex-dump:

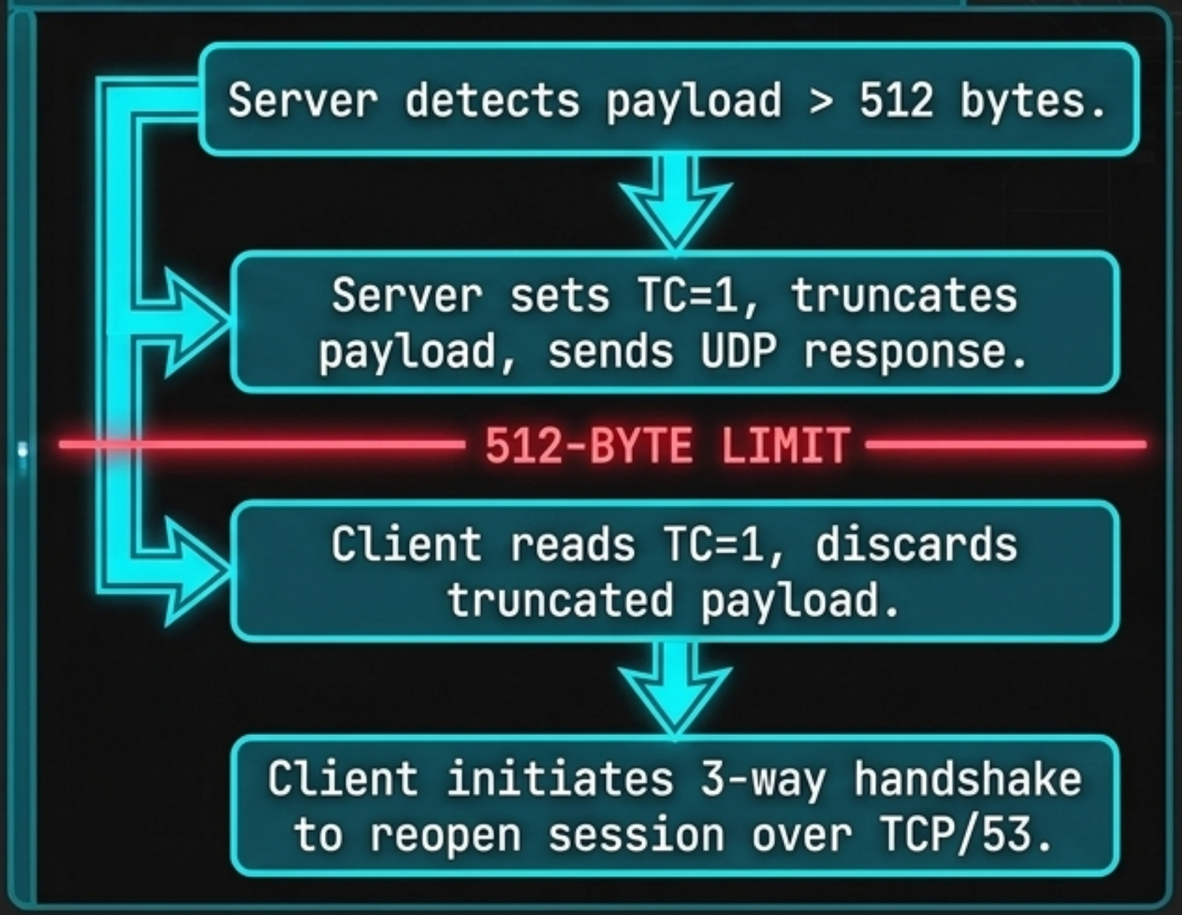
| Transaction ID | |
|---|---|
| Flags | |

| QR | Opcode | AA | TC | RD | RA | Z | RCODE |
|---|---|---|---|---|---|---|---|

15 14 13 12 11 10 9 8 | 7 6 5 4 3 2 1 0

| QDCOUNT | ANCOUNT |
|---|---|
| 16 bits | 16 bits |
| NSCOUNT | ARCOUNT |
| 16 bits | 16 bits |

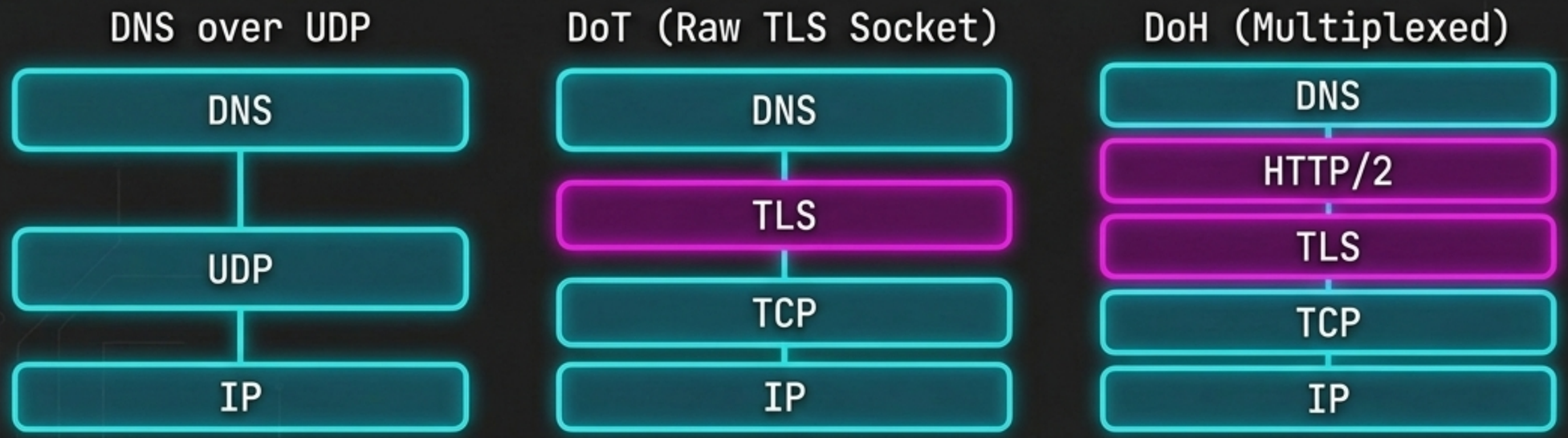
## Logical Flowchart: UDP Truncation




Gemini Notebook

# The Encryption Era: DoH (RFC 8484)

DoH maps a DNS query to an HTTP exchange:
- GET requests use base64url encoding (without padding) in the URI: ?dns=AAABAAAB...
- Media type: application/dns-message
- Leverages HTTP/2 compression and multiplexing.




ARCHITECTURAL TENSION: Encrypting the last mile prevents local ISP snooping/hijacking, but centralizes global resolution data into a handful of mega-resolvers (e.g., 1.1.1.1, 8.8.8.8), altering the decentralized design of the original internet.

Gemini Notebook

# Core Record Semantics & Negative Caching

ZONE FILE SNIPPET (SOA RECORD)

```
@       IN  SOA ns1.example.com.
                hostmaster.example.com.
                ( 2023102401; SERIAL
                  7200      ; REFRESH
                  600       ; RETRY
                  3600000   ; EXPIRE
                  300 )     ; MINIMUM
```

**MNAME**: Primary source of truth.

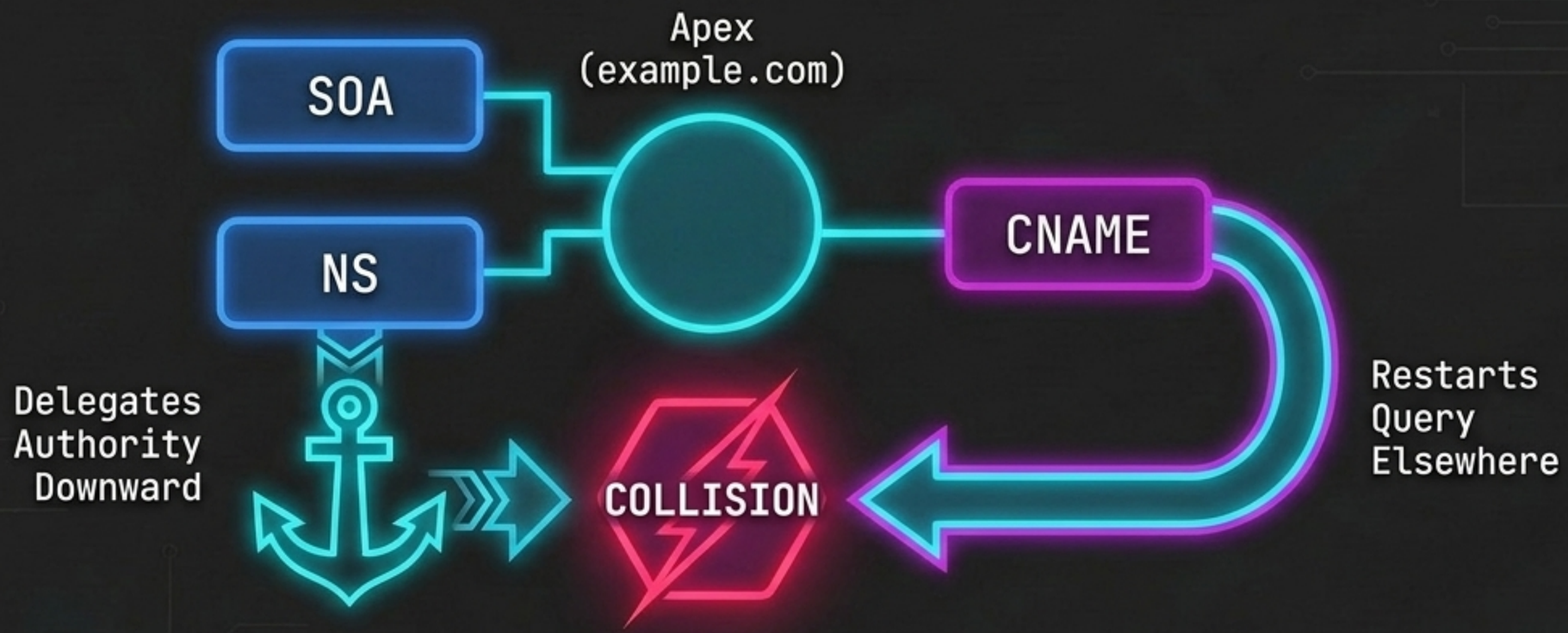
**RNAME**: Responsible admin mailbox.

**MINIMUM**: Establishes the baseline TTL for negative caching (NXDOMAIN / empty NOERROR).
A resolver caches the non-existence of a record for this duration.

Gemini Notebook

# The Zone Apex Collision (RFC 1034)

RFC 1034 dictates that if a CNAME RR is present at a node, no other data should be present.

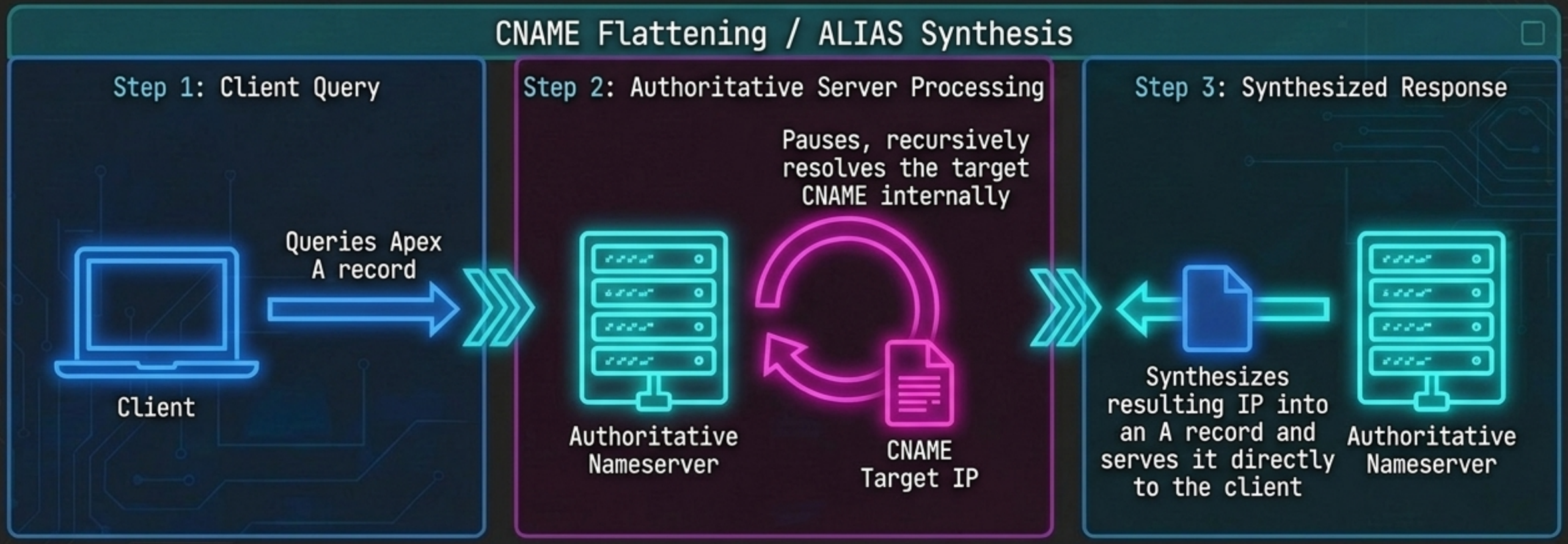



The Paradox: A domain's apex inherently requires SOA and NS records to establish authority. Placing a CNAME here forces the resolver to restart the query at the target, completely ignoring the NS delegation. The zone structurally ceases to exist.

Gemini Notebook

# Synthetic Workarounds: CNAME Flattening

**Mechanism**: Resolves the apex collision by shifting the indirection from the client's resolver to the authoritative server. The client never sees the CNAME; it receives a perfectly compliant, dynamically generated A/AAAA record.




Step 1: Client Query

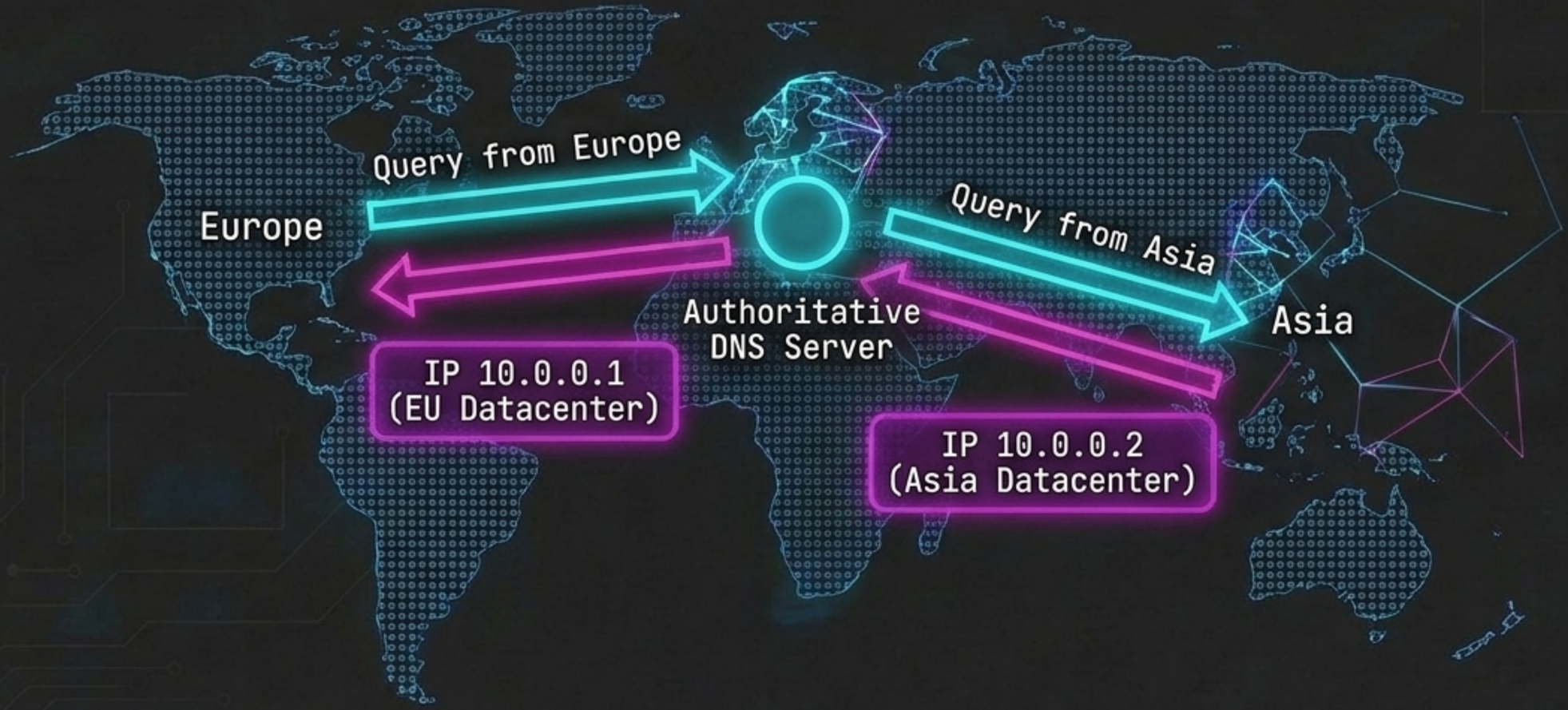
Step 2: Authoritative Server Processing

Step 3: Synthesized Response

Gemini Notebook

# Unicast GeoDNS Steering Mechanics

**Core Concept**: Steering traffic by mapping the incoming recursive resolver's IP address to an ASN/GeoIP database. The authoritative server calculates the lowest-latency datacenter and serves a customized A/AAAA record.

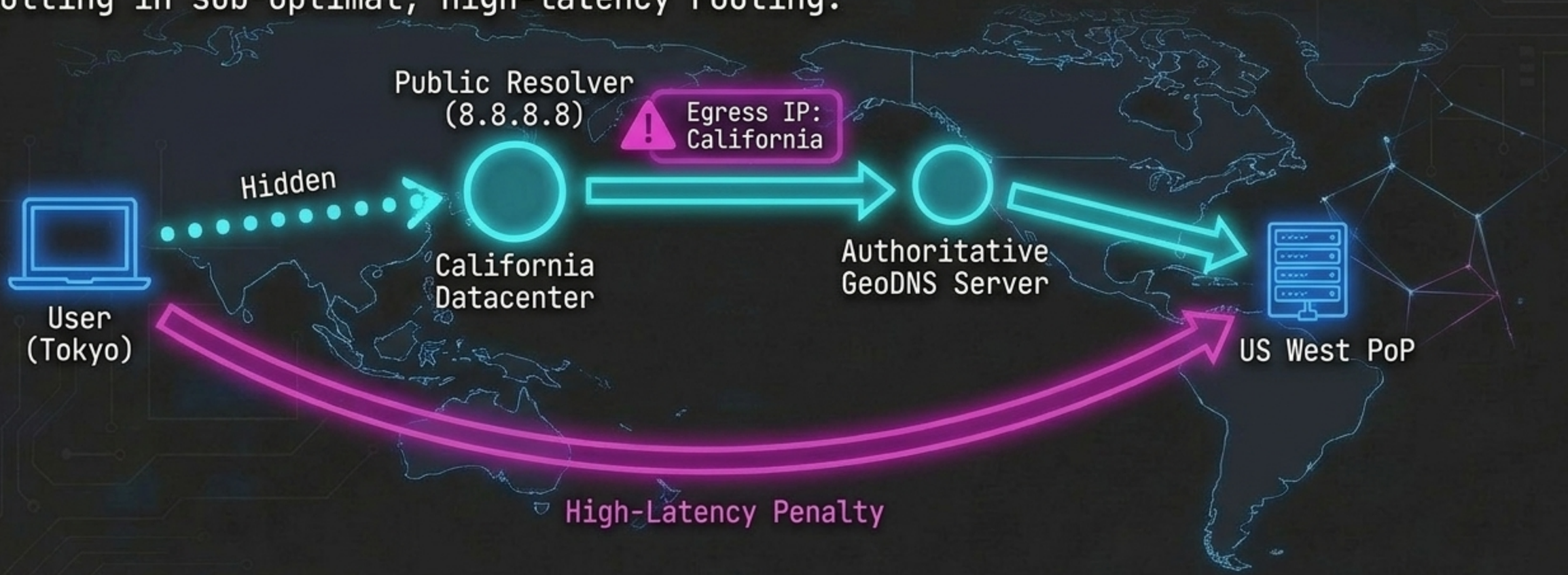



Gemini Notebook

# The Public Resolver Blindspot

**The Flaw**: The IP address seen by the authoritative nameserver belongs to the recursive resolver, not the end user.

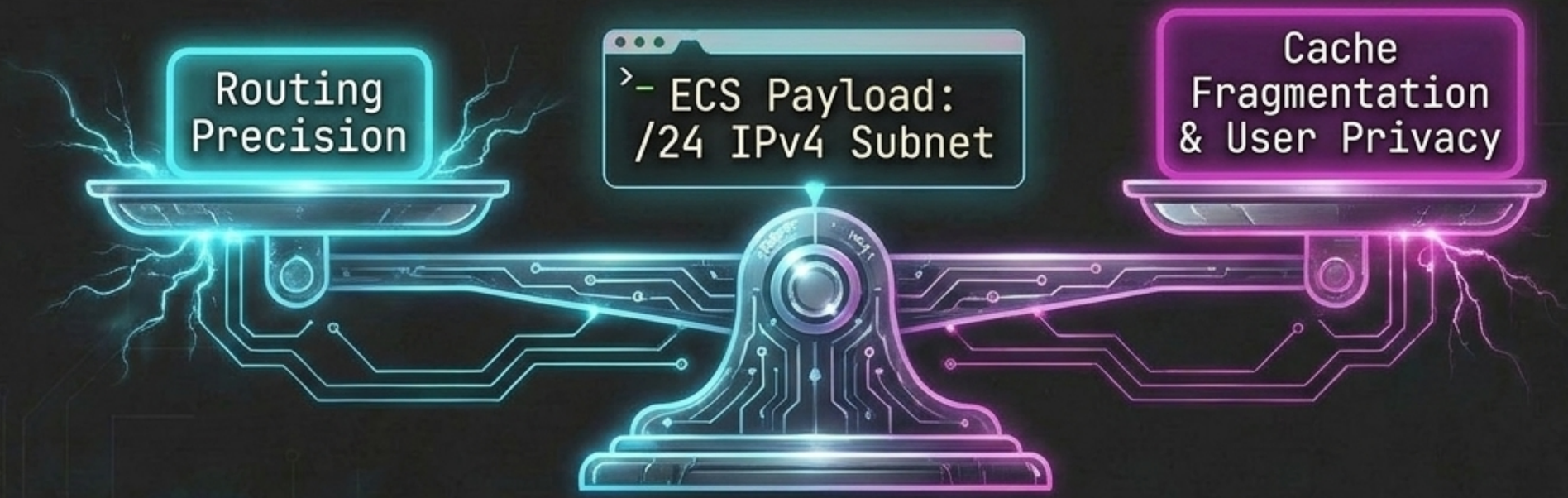
**Impact**: As users increasingly rely on centralized mega-resolvers (Google DNS, Cloudflare), regional origin context is stripped, breaking GeoDNS accuracy and resulting in sub-optimal, high-latency routing.




Gemini Notebook

# EDNS Client Subnet (ECS) - RFC 7871

Mechanism: The recursive resolver passes a truncated client IP prefix in the EDNS0 extension data, restoring geographic context for the authoritative server.



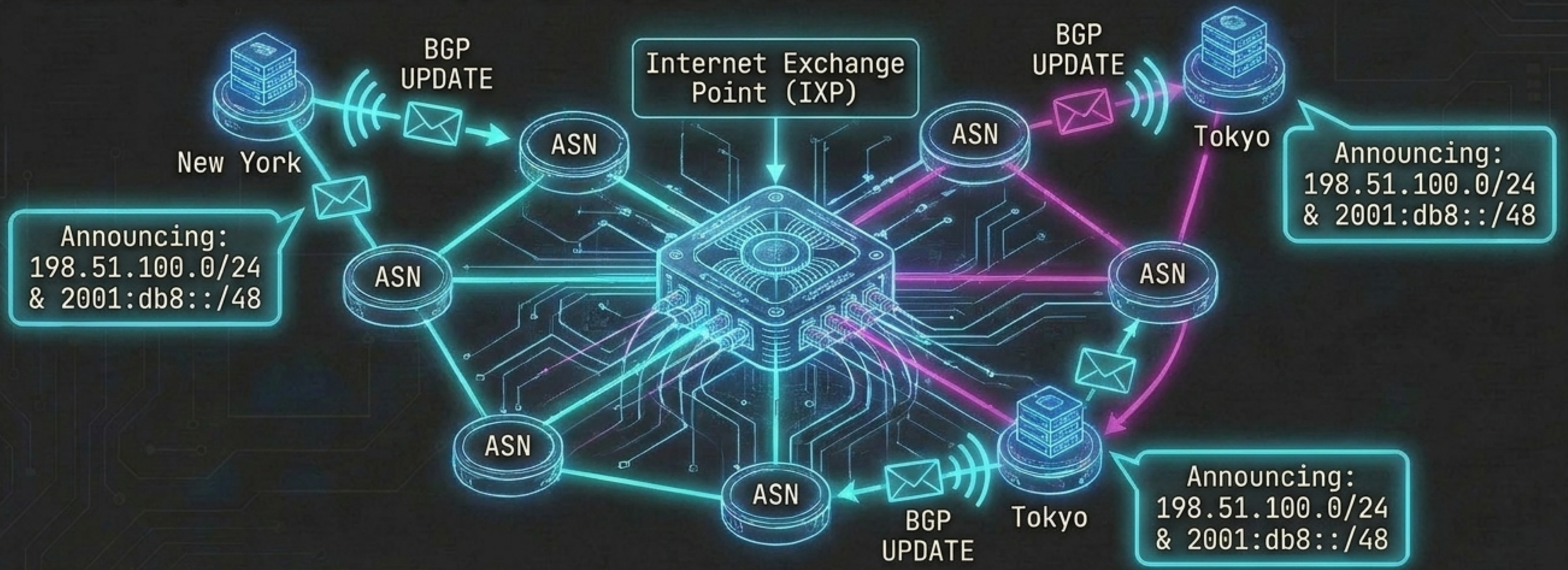

- Cache Fragmentation: Each unique ECS prefix effectively fragments the recursive resolver's cache, exponentially lowering cache hit rates.
- Privacy: Exposes localized client subnet data to upstream authoritative operators.

Gemini Notebook

# BGP-4 Fundamentals for the Edge (RFC 4271)

**Core Concept**: Anycast discards DNS-level steering. Instead, multiple distinct edge locations advertise identical IP space to the global routing table using eBGP.

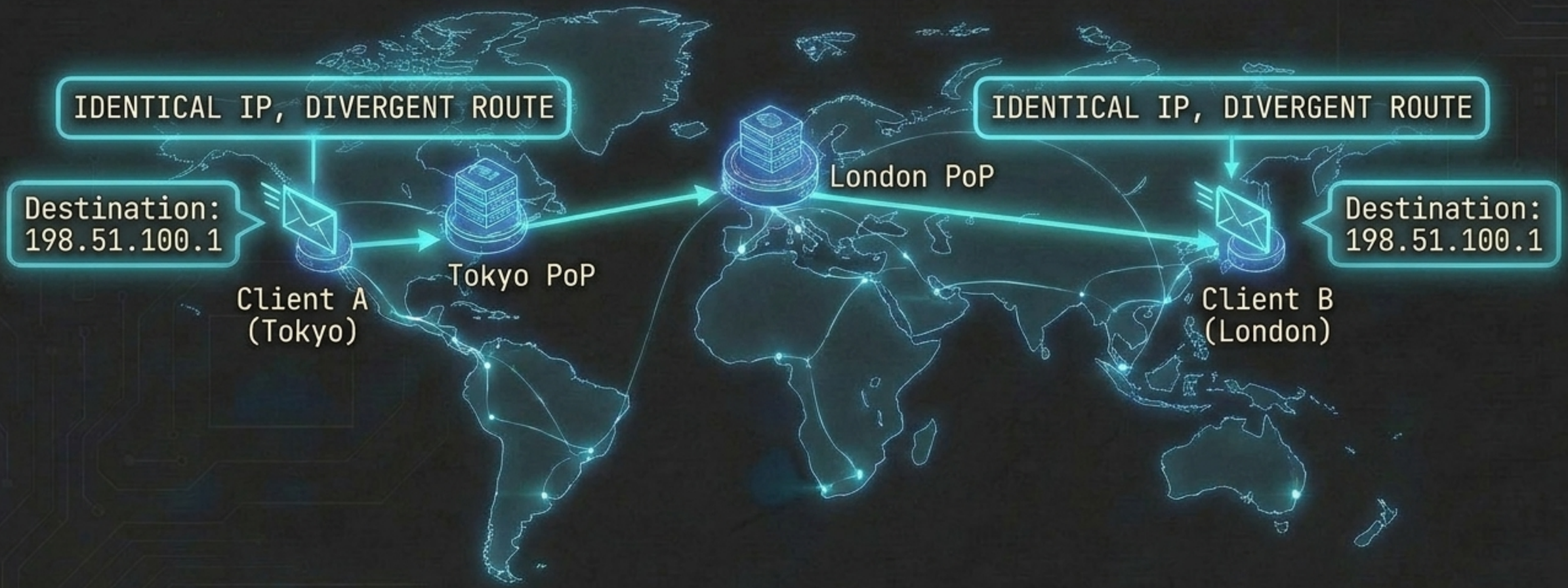
**Constraint**: To propagate globally across the default-free zone, Anycast requires minimum prefix sizes of /24 (IPv4) and /48 (IPv6).




Gemini Notebook

# The Anycast Illusion: One IP, Many PoPs

**Mechanism**: Moving load balancing from the DNS layer to the BGP layer. End users do not receive different IPs based on location; they all receive the same IP, and the transit network natively routes the packet to the topologically nearest Point of Presence (PoP).




Gemini Notebook

# BGP Path Selection vs. Physical Reality

**The Reality Check**: BGP is fundamentally unaware of geographic distance or fiber latency. It routes strictly on policy and path vectors.

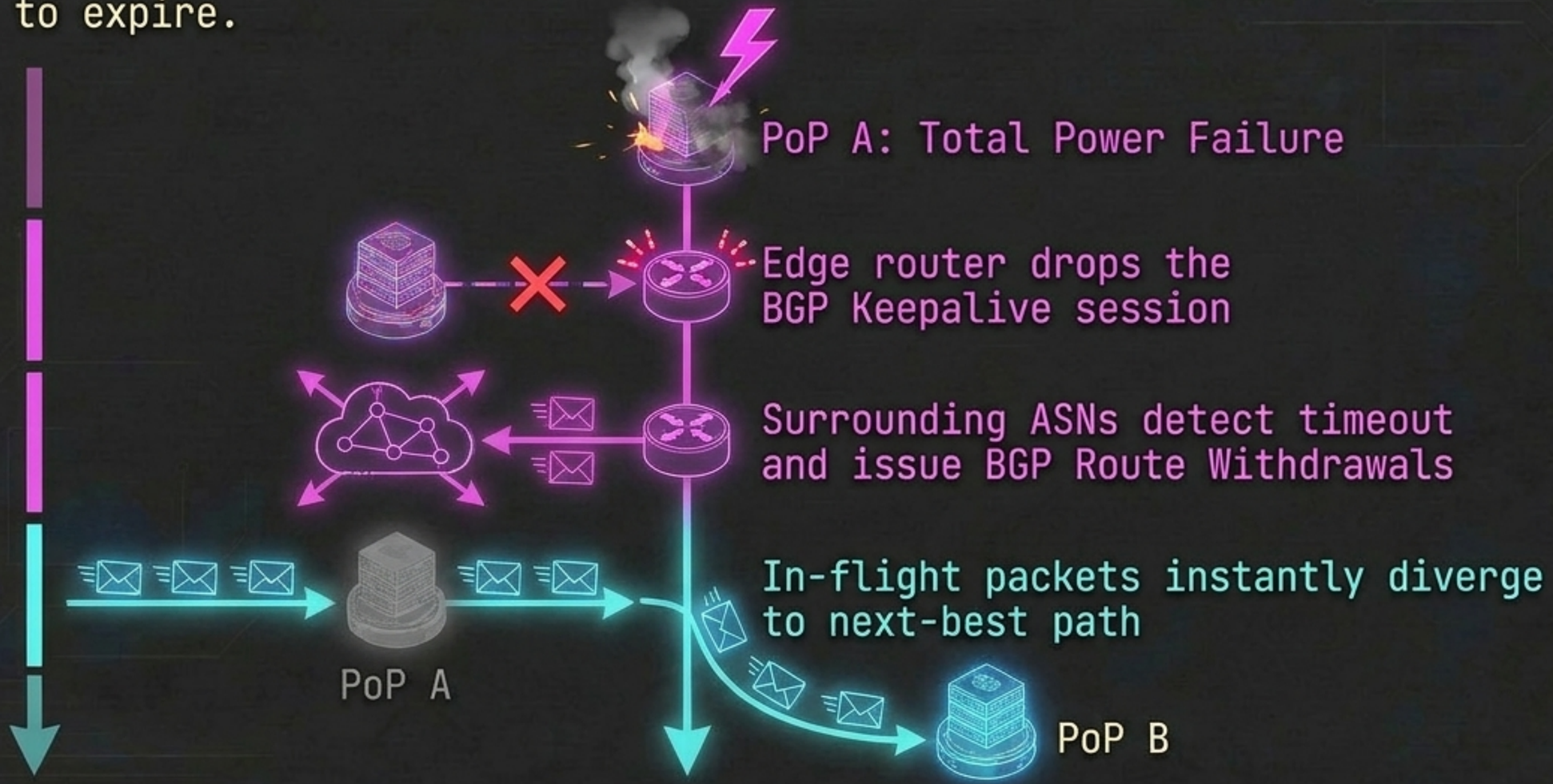
**Metrics**: Path selection prioritizes LOCAL_PREF (economic policy), then AS_PATH (hop count), and MED (exit point).




Gemini Notebook

# Outages and PoP Withdrawal

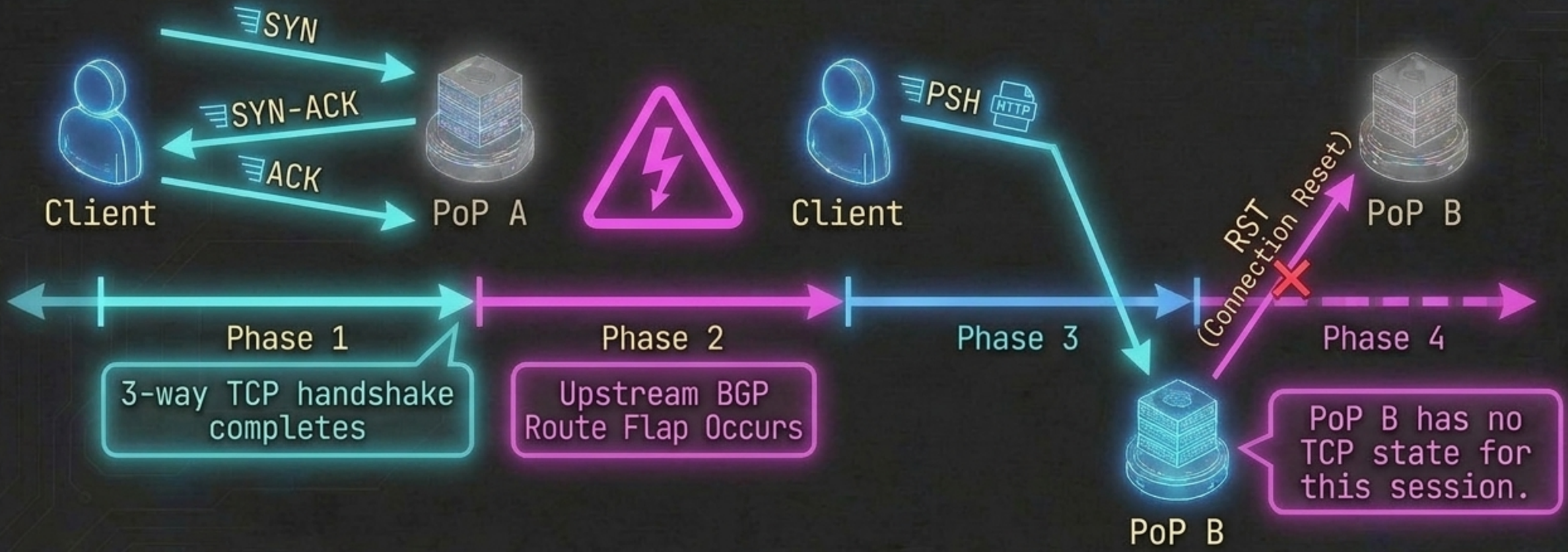
The Failover Advantage: Unlike DNS failover, which is crippled by caching and TTL delays, Anycast failover occurs at the network layer. If a PoP dies, the route is withdrawn, and global traffic instantly reroutes without waiting for a single client TTL to expire.




Gemini Notebook

# The Route Churn Threat

The Vulnerability: Anycast is stateless; TCP/QUIC are highly stateful. If upstream transit volatility causes a mid-flight packet to hit a different datacenter, the connection breaks.

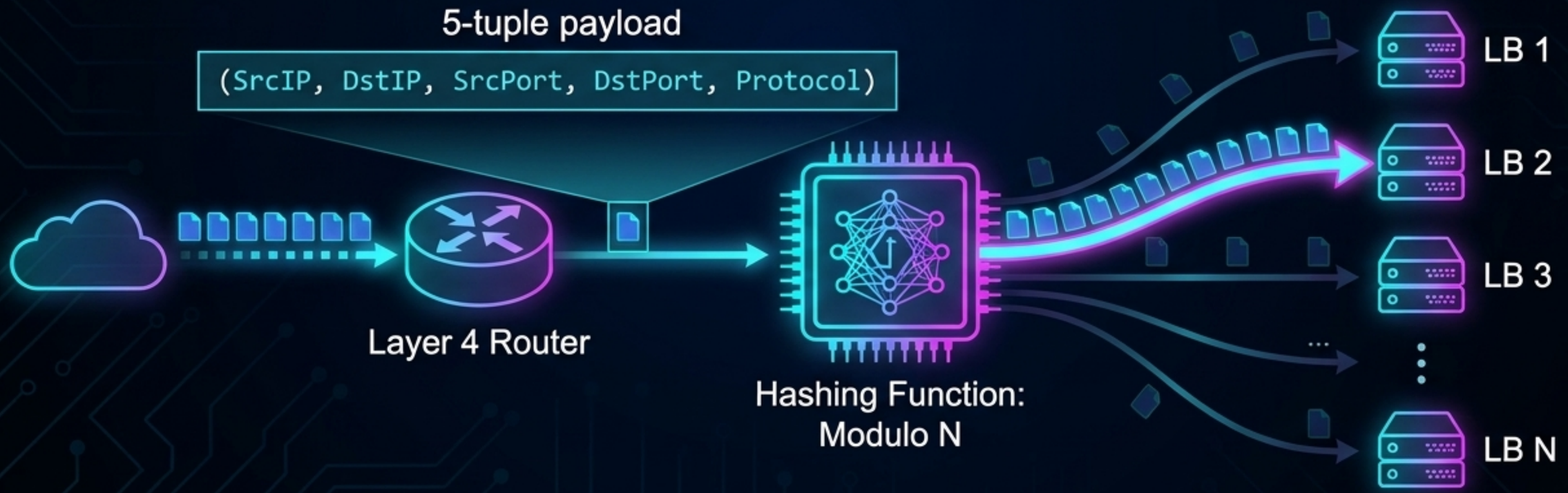



Gemini Notebook

# ECMP 5-Tuple Hashing

**Mechanism:** Equal-Cost Multi-Path (ECMP) routing hashes the 5-tuple of every incoming packet: `id(packet) mod N`. This deterministically pins all packets belonging to a single flow to the **exact same** Layer 4 load balancer.

**The Problem:** If a single load balancer dies, N becomes N-1. The hash `id(packet) mod N-1` remaps nearly all existing connections to different servers, resetting millions of active sessions.

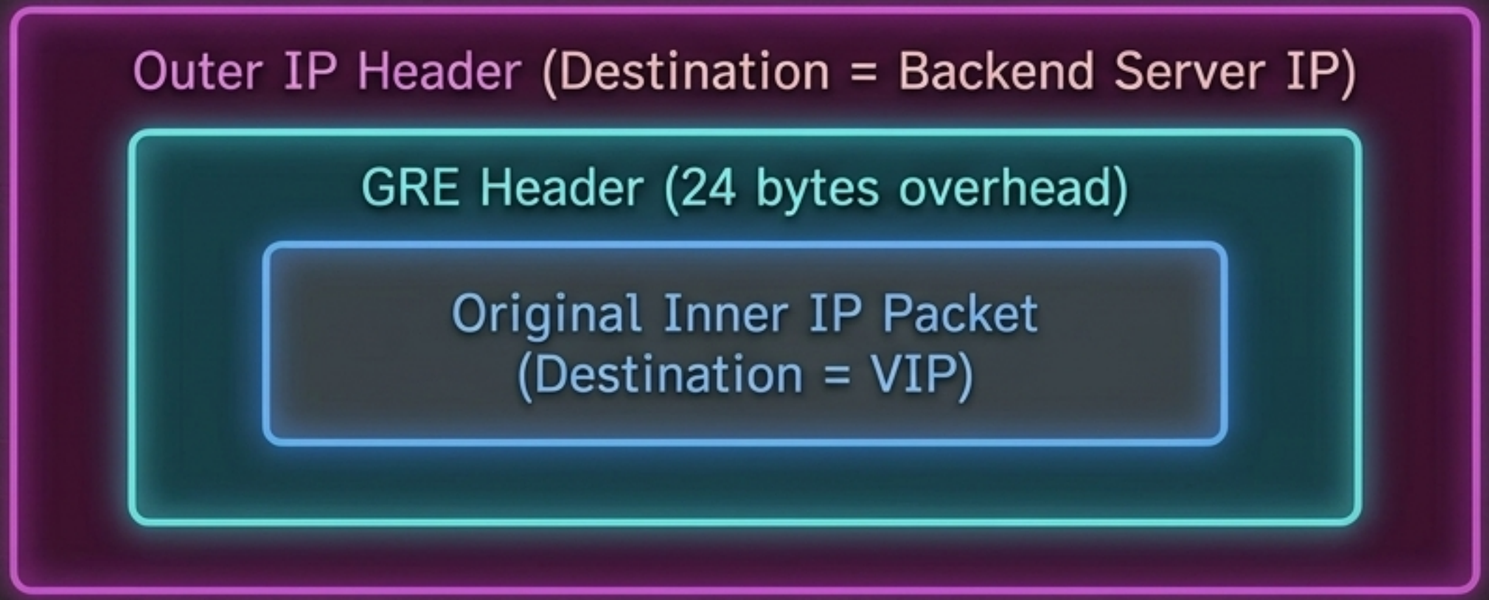



Gemini Notebook

# Stateful Forwarding, Maglev & GRE

**Consistent Hashing:** Solves the N-1 threat by ensuring that when a backend fails, only the flows connected to the dead machine are remapped; the rest remain pinned.

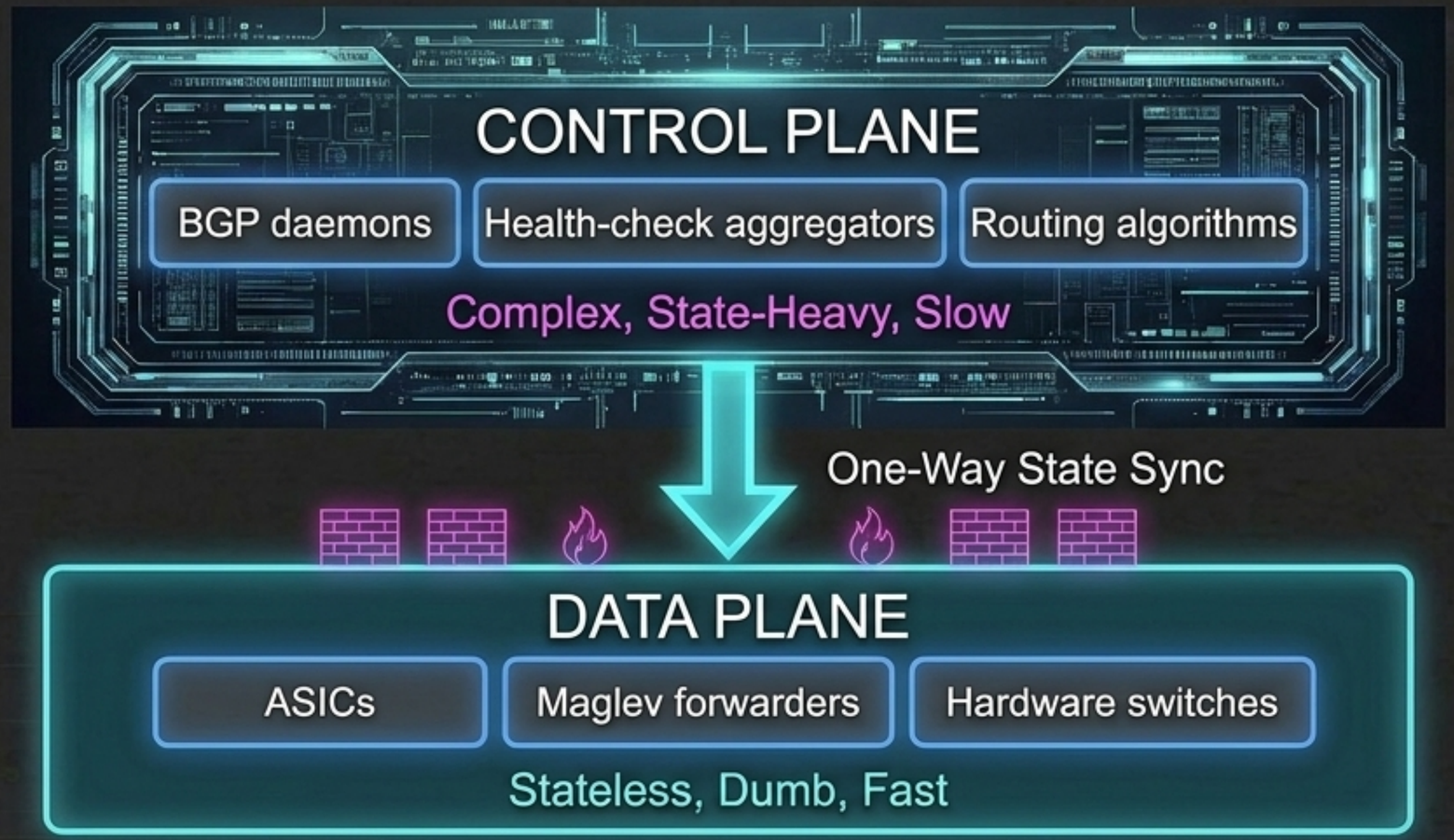
**Encapsulation vs DSR:** Traditional Direct Server Response (DSR) alters MAC addresses at Layer 2, trapping all load balancers in a single broadcast domain. Encapsulating the packet in IP+GRE allows the backend to be physically anywhere, at the cost of MTU overhead.




Gemini Notebook

# Isolation: Control Plane vs. Data Plane

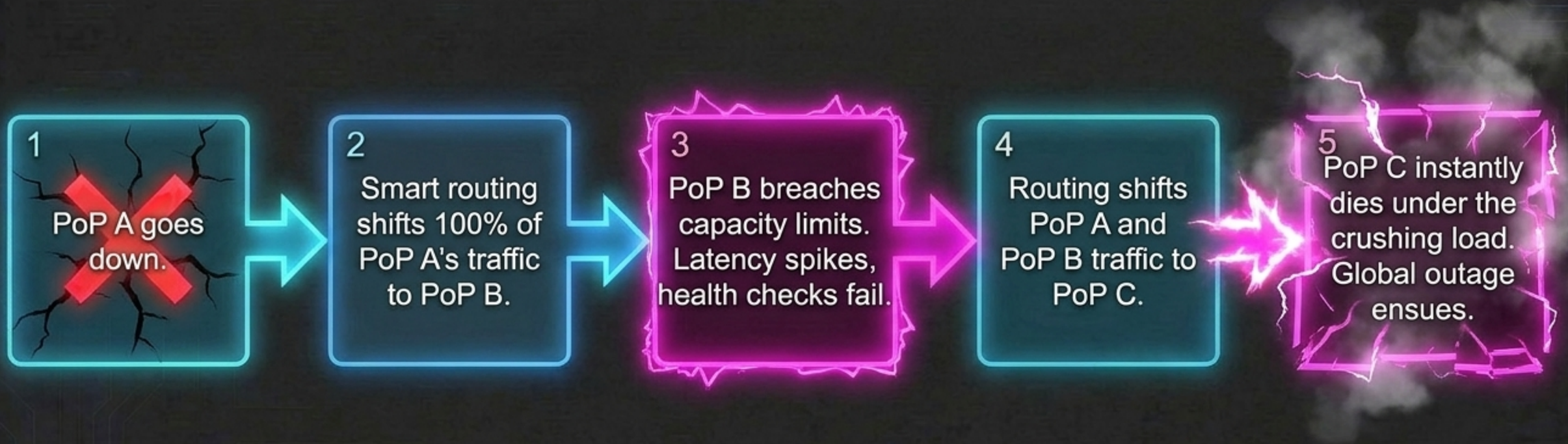
**The SRE Mandate:** The systems that compute routing logic must be strictly decoupled from the systems pushing bits on the wire. If the Control Plane suffers a catastrophic memory leak or database corruption, the Data Plane must continue forwarding traffic blindly using its last known state.




Gemini Notebook

# The Dynamic Reconfiguration Cascade

**The Danger:** Highly responsive control planes can amplify localized outages into global cascades. Shifting traffic triggers overload logic in healthy PoPs, transforming a partial degradation into a total platform collapse.

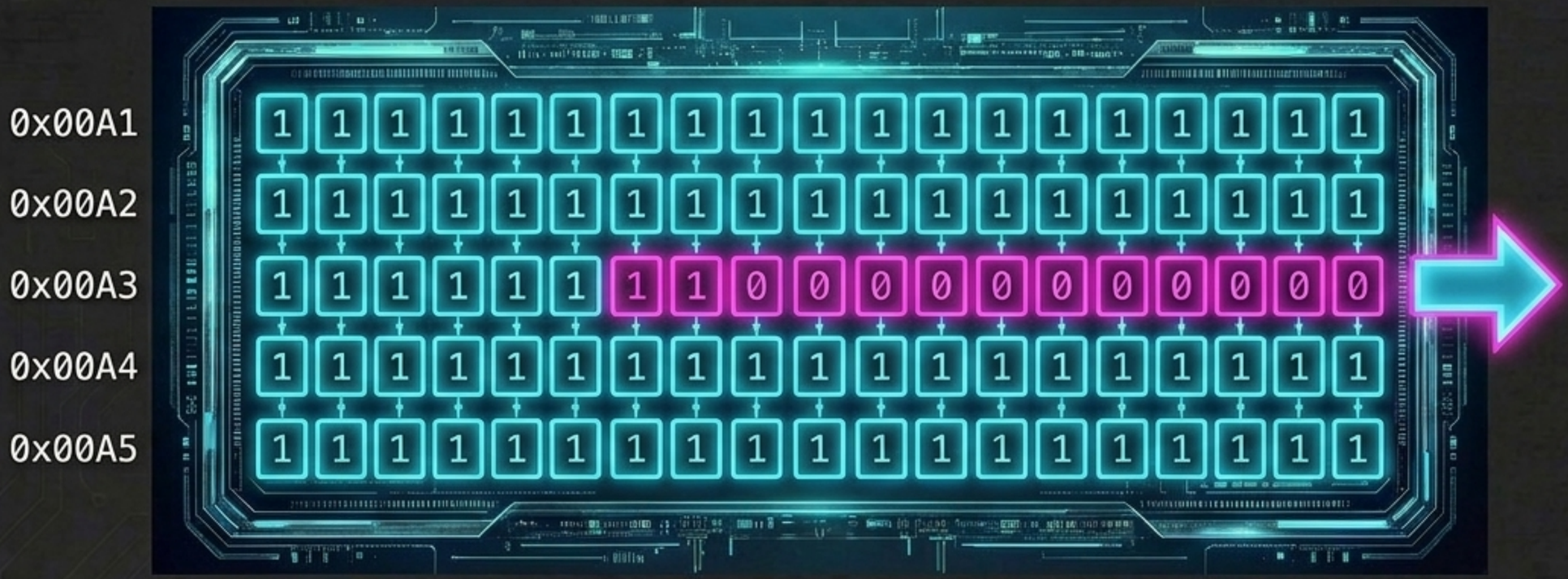



Gemini Notebook

# Constant-Work Health Checking

**Design Philosophy:** A system's computational cost must be identical during 100% uptime and 100% downtime.

**Implementation:** By evaluating route health through static memory structures (bitmaps), checking the status of 10,000 servers takes exactly the same CPU cycles in peacetime as it does during a catastrophic datacenter fire. Emergency states do not generate unpredictable CPU spikes.




**O(1) Evaluation:** Scanning memory blocks costs fixed CPU cycles regardless of health state.

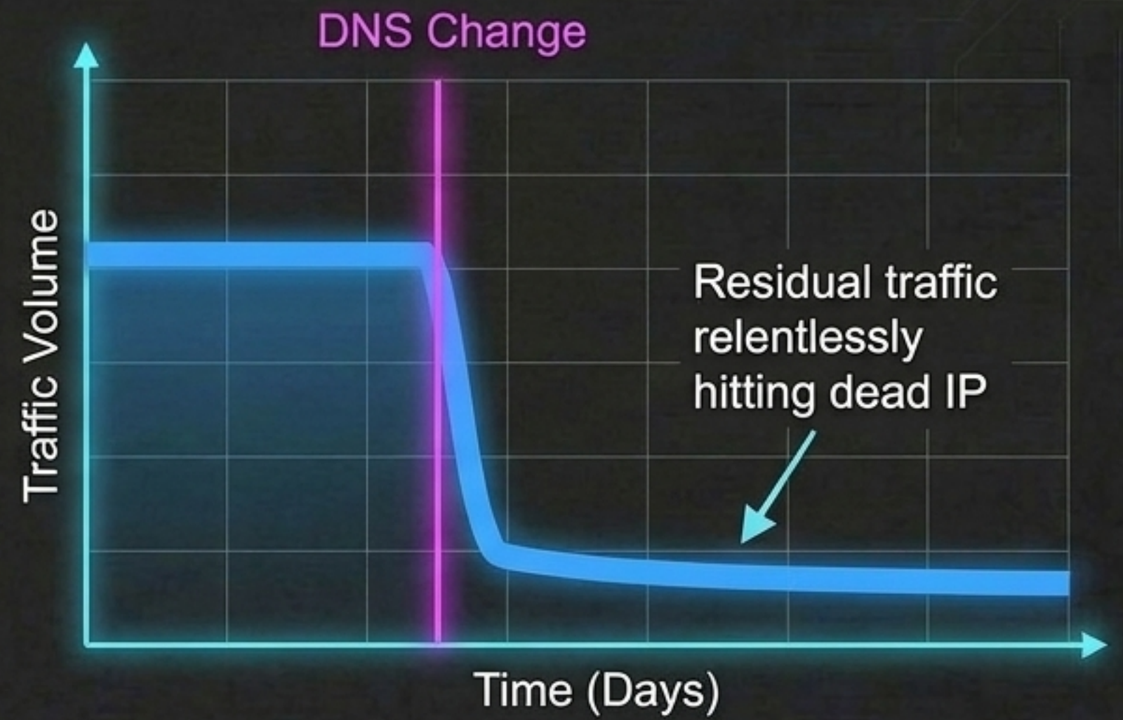
Gemini Notebook

# Failure Mode 1: The Stubborn Client

**The Reality:** You do not control your DNS TTLs; the client does. Browsers, ISPs, and default JVM configurations notoriously ignore low TTLs to mask latency.

**Impact:** Relying solely on Unicast DNS for emergency failover guarantees a long tail of dropped traffic as rogue client connection pools fiercely hold onto stale IP addresses.

```
# java.security configuration file
networkaddress.cache.negative.ttl=10
networkaddress.cache.ttl=-1   // Cache forever
...
```

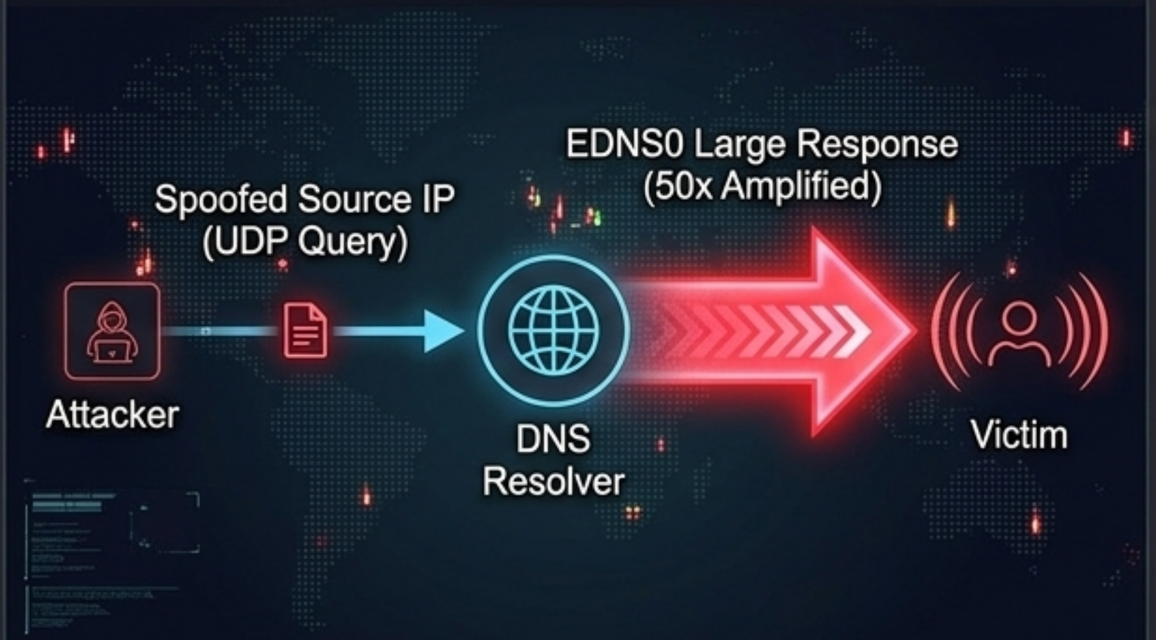



Gemini Notebook

# Failure Mode 2: Protocol Weaponization

## DNS Amplification

UDP's stateless nature allows attackers to spoof source IPs. Large EDNS0 replies are weaponized as volumetric firehoses against victims.




## BGP Hijacking

Because BGP fundamentally operates on trust, rogue ASNs can steal prefixes. The defense is **RPKI** (Resource Public Key Infrastructure) generating ROAs to cryptographically validate route announcements.




Gemini Notebook

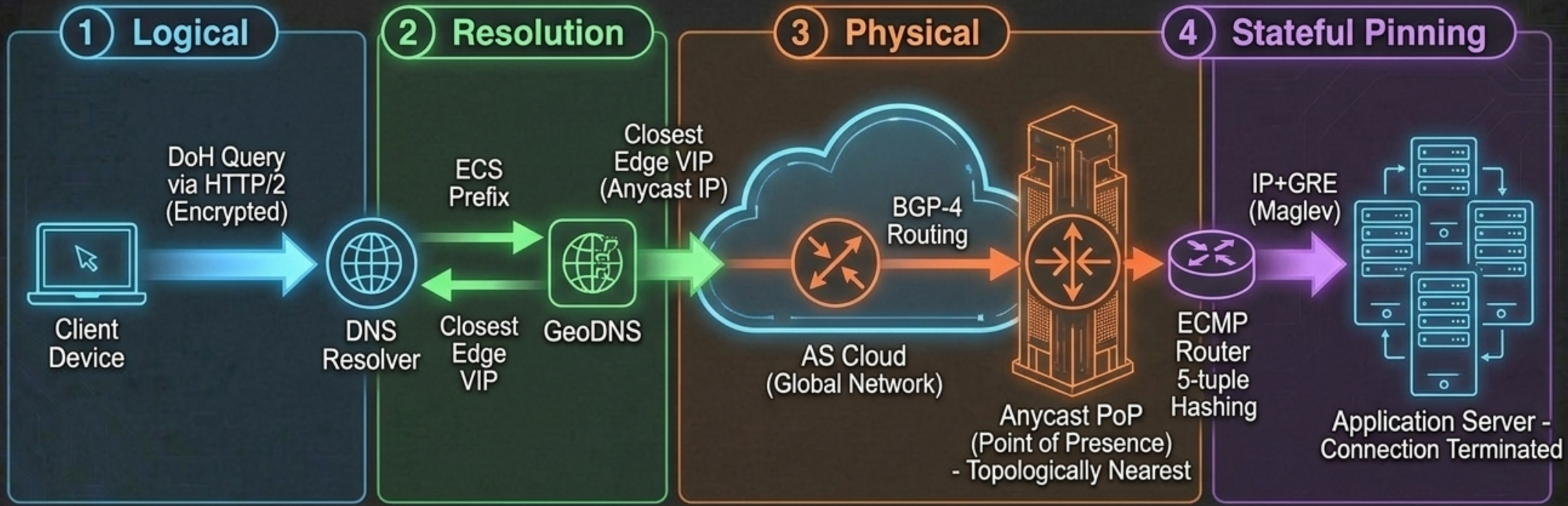
# The Global Server Load Balancing (GSLB) Matrix

| | Anycast Network Edge | Unicast GeoDNS |
|---|---|---|
| **Wins** | • Unsurpassed DDoS mitigation (sinkholes localized attacks)<br>• Instant BGP failover<br>• Zero reliance on client TTL compliance | • Granular traffic shaping and weighting<br>• Targeted canary rollouts<br>• Capacity-aware load shedding |
| **Fails** | • Cannot easily weight traffic (e.g., send exactly 10% to PoP A)<br>• Stateful connections vulnerable to route flaps | • Slower failover bound by stubborn TTL caches<br>• Resolver masking breaks geo-accuracy |

Gemini Notebook

# Synthesis: The Packet's Journey to the Edge

**Core Takeaway:** Modern web scale requires mastering the symbiosis between logical Application Layer routing (DNS) and physical Network Layer constraints (BGP).




Gemini Notebook