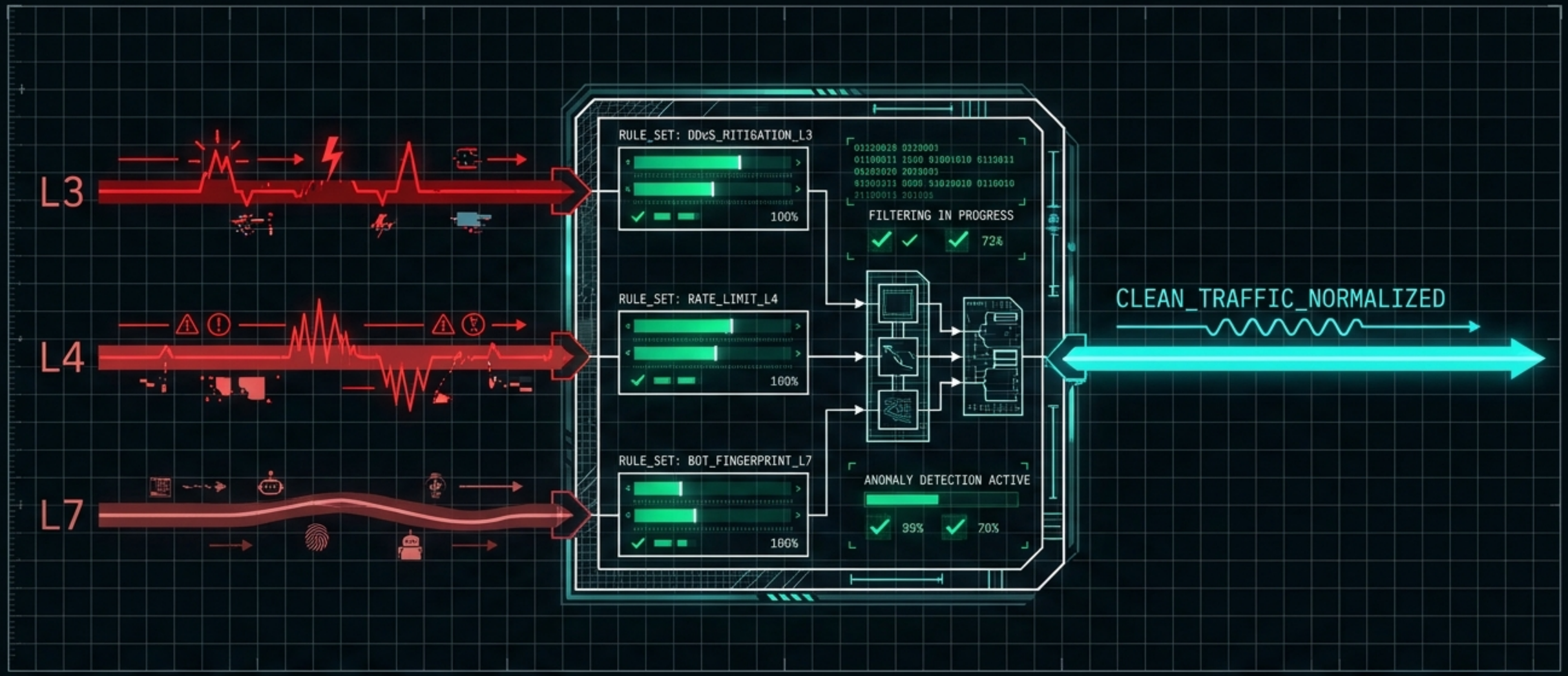


Web Architecture — S02E03: WAF Internals, L3/L4/L7 DDoS Mitigation, Bot Fingerprinting & Edge Rate Limiting

CONFIDENTIAL ENGINEERING WHITEPAPER // FOR INTERNAL DISTRIBUTION ONLY

Gemini Notebook

Multi-Layer Attack Taxonomy

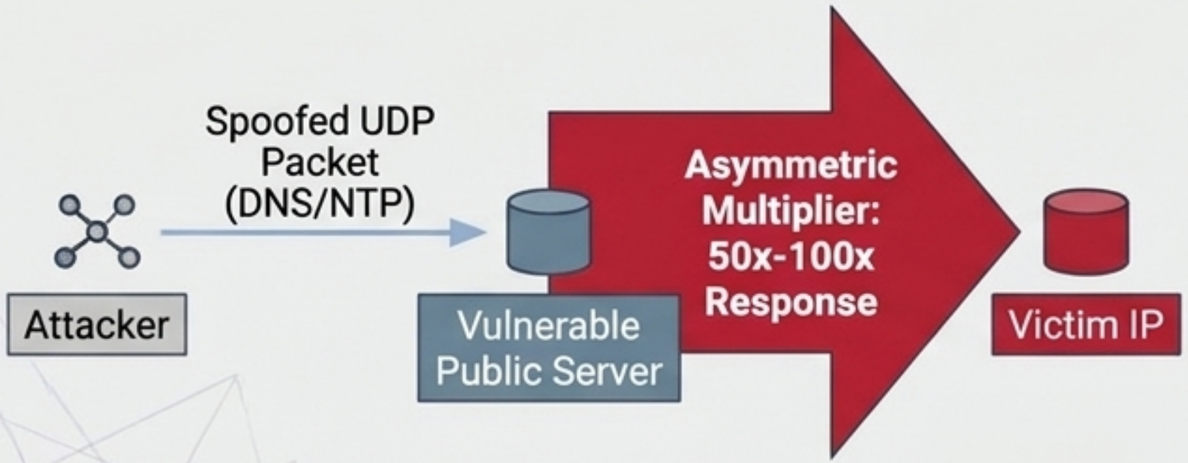
The attack surface spans volumetric congestion, state exhaustion, and application logic

Layer Focus	Attack Category	Primary Vector	Bandwidth Cost to Attacker	Target Resource	Detection Difficulty
Layer 3 - Network	Volumetric	ICMP/IPsec/UDP	Very Low	Network Capacity	Easy (Static Thresholds)
Layer 4 - Transport	Protocol	TCP SYN / ACK	Low	Load Balancer/Firewall State Tables	Moderate
Layer 7 - Application	Application	HTTP GET/POST	High (Requires full TCP connection)	CPU/DB/API Compute	High (Blends with legitimate traffic)

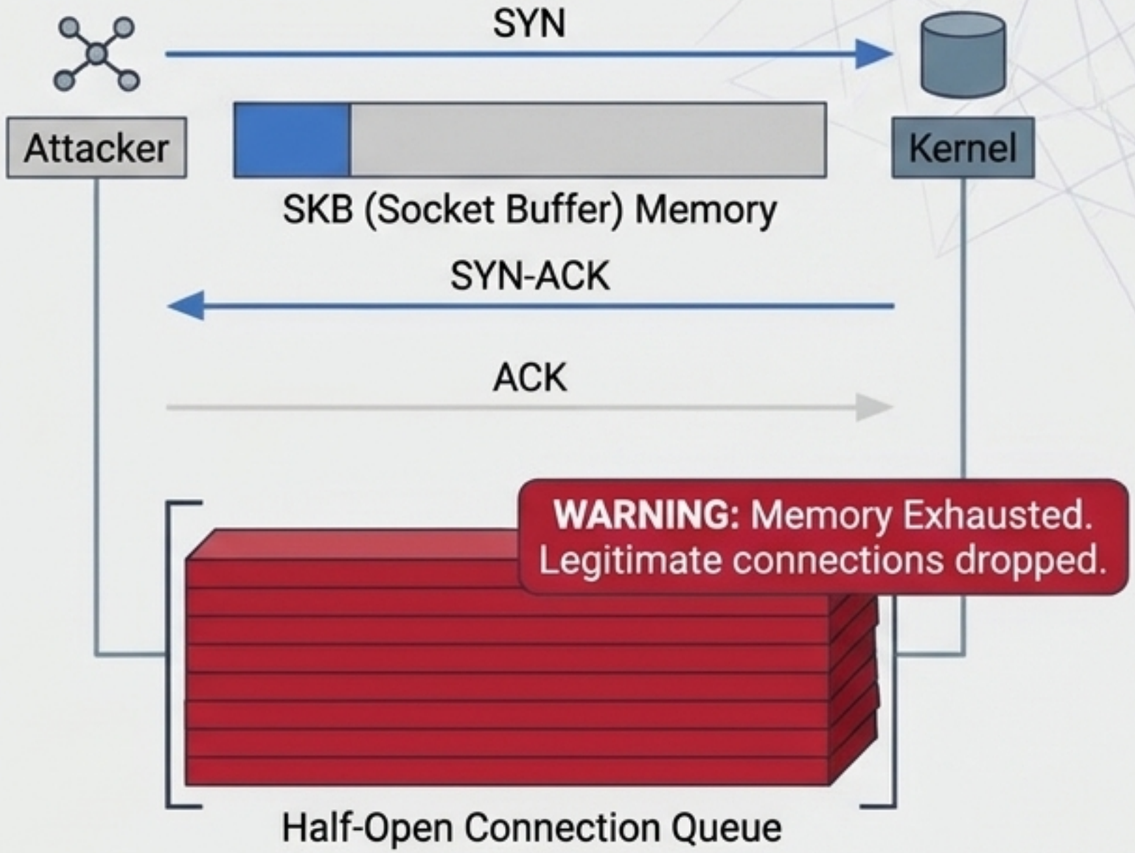
Gemini Notebook

Network-layer attacks leverage bandwidth amplification and kernel state exhaustion

UDP Reflection/Amplification

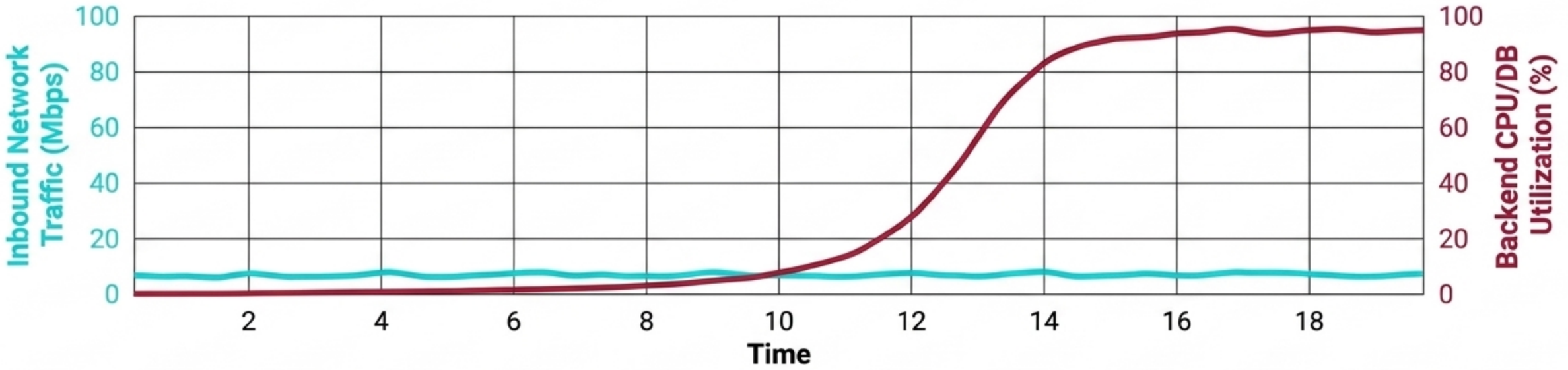


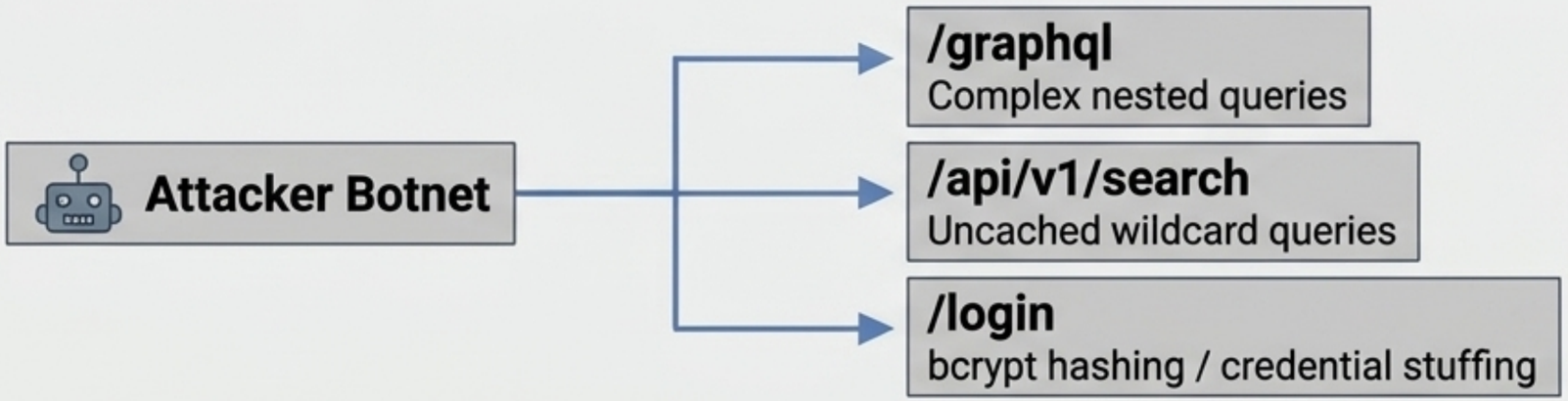
TCP SYN Flood



Gemini Notebook

Layer-7 HTTP floods exploit resource asymmetry to crash backend databases

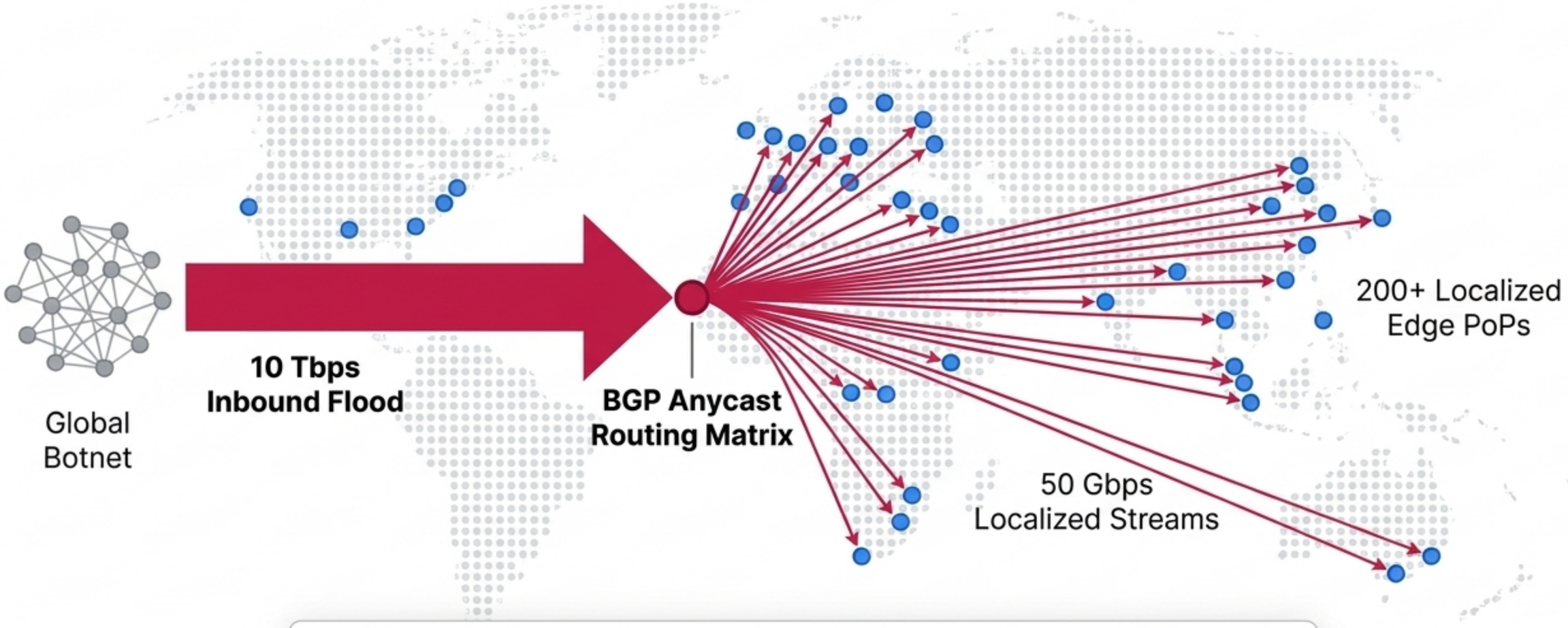




Because a single HTTP request requires expensive DB queries and API calls to fulfill, attackers can crash the origin server without triggering volumetric network alarms.

Gemini Notebook

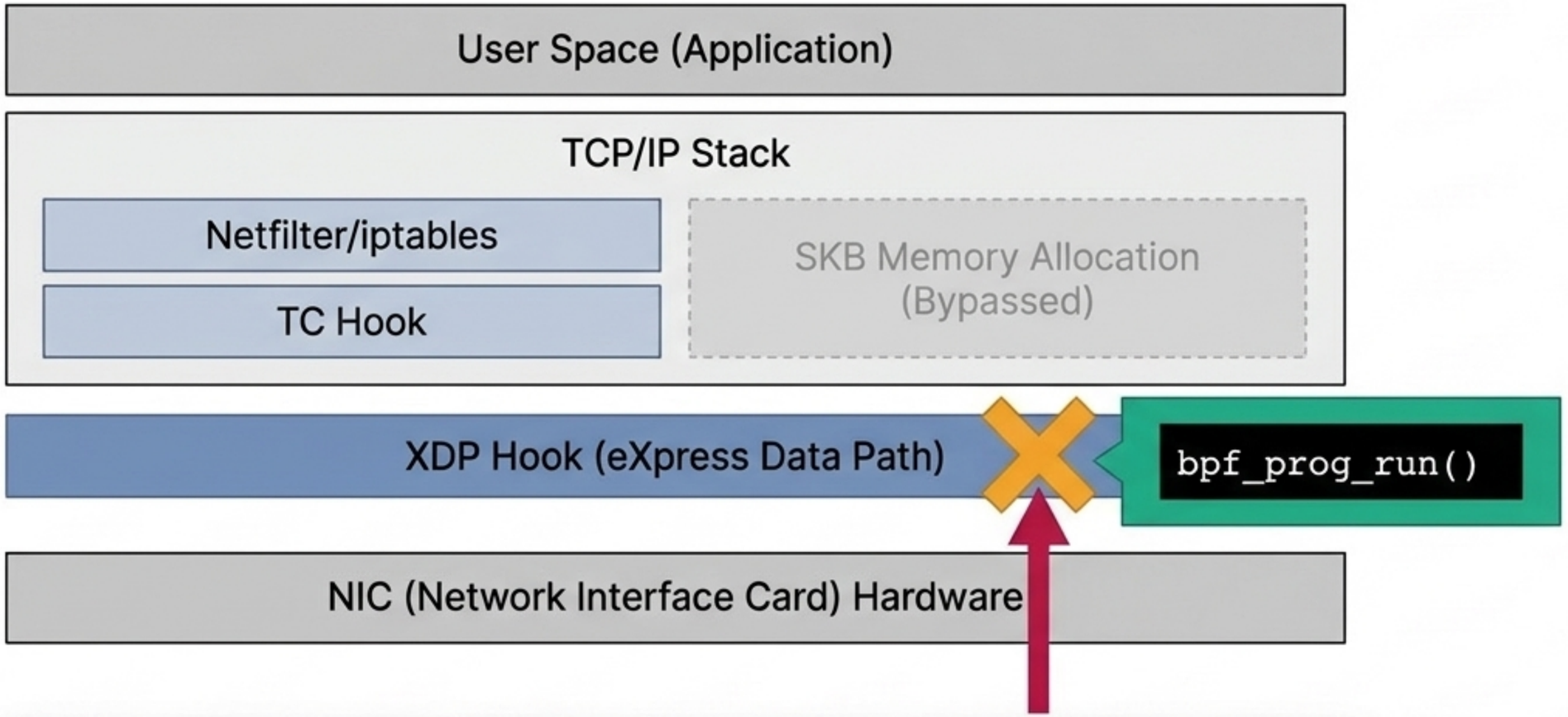
BGP Anycast routing diffuses terabit-scale volumetric floods across global edge PoPs



Mitigation is achieved by consuming the attack with overwhelming localized network capacity before it reaches the origin.

Gemini Notebook

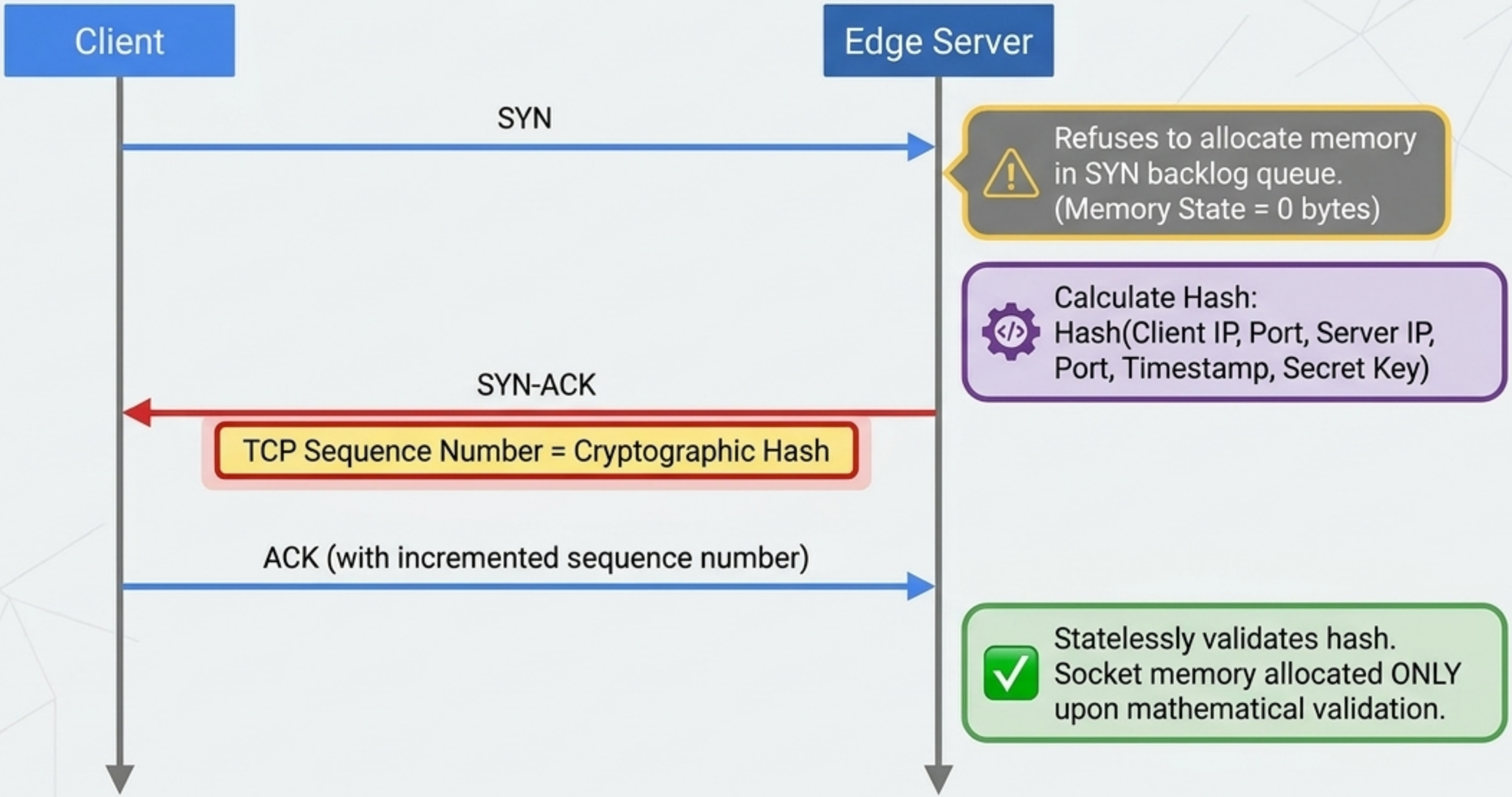
eBPF and XDP drop malicious packets at the network driver level to preserve memory



Malicious packets are dropped at the lowest software level before costly SKB memory allocation can occur, saving CPU cycles during a SYN flood.

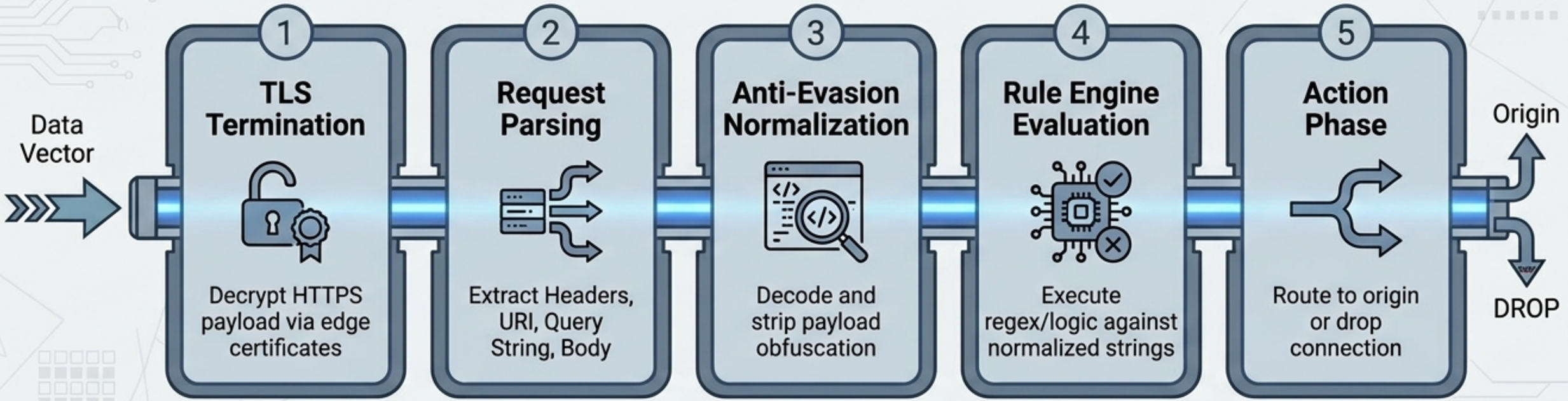
Gemini Notebook

Stateless TCP SYN cookies defer memory allocation via cryptographic hashing



Gemini Notebook

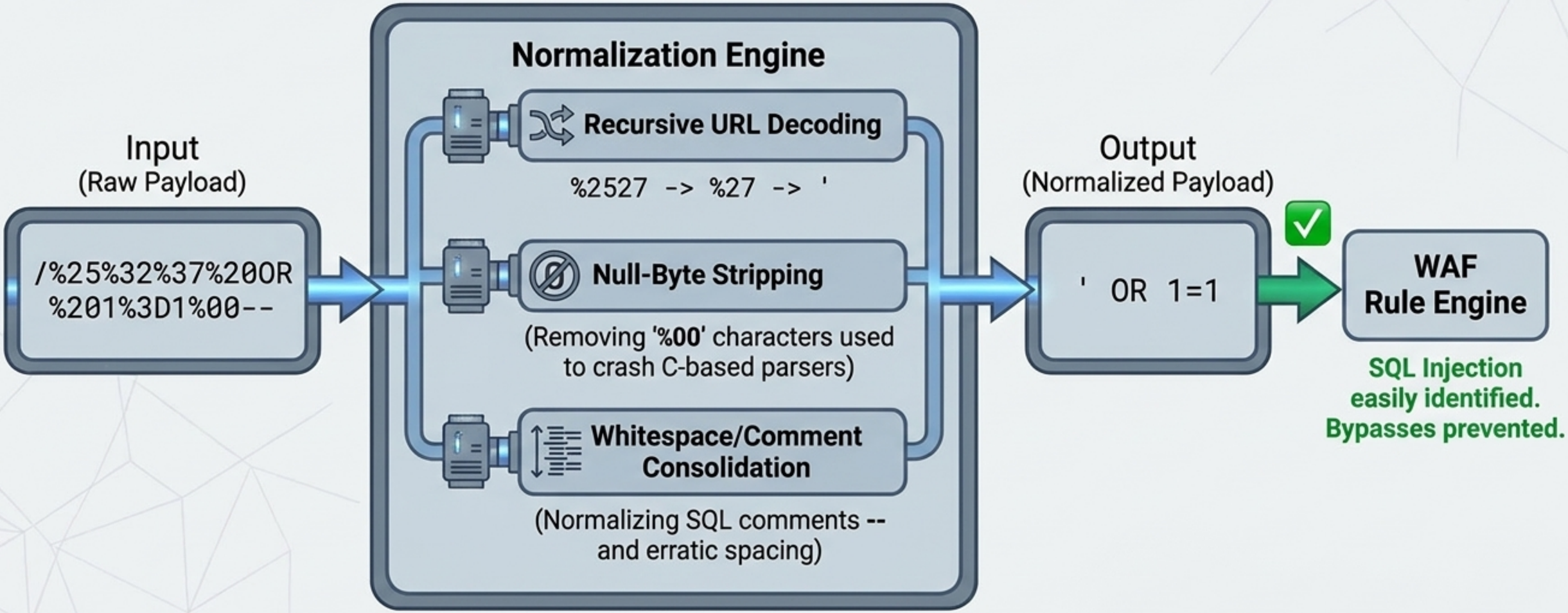
Layer 7 inspection executes a deterministic, multi-stage request pipeline



The WAF operates as a **reverse-proxy shield**, preventing direct client exposure to the origin server by forcing all traffic through this sequential execution gauntlet.

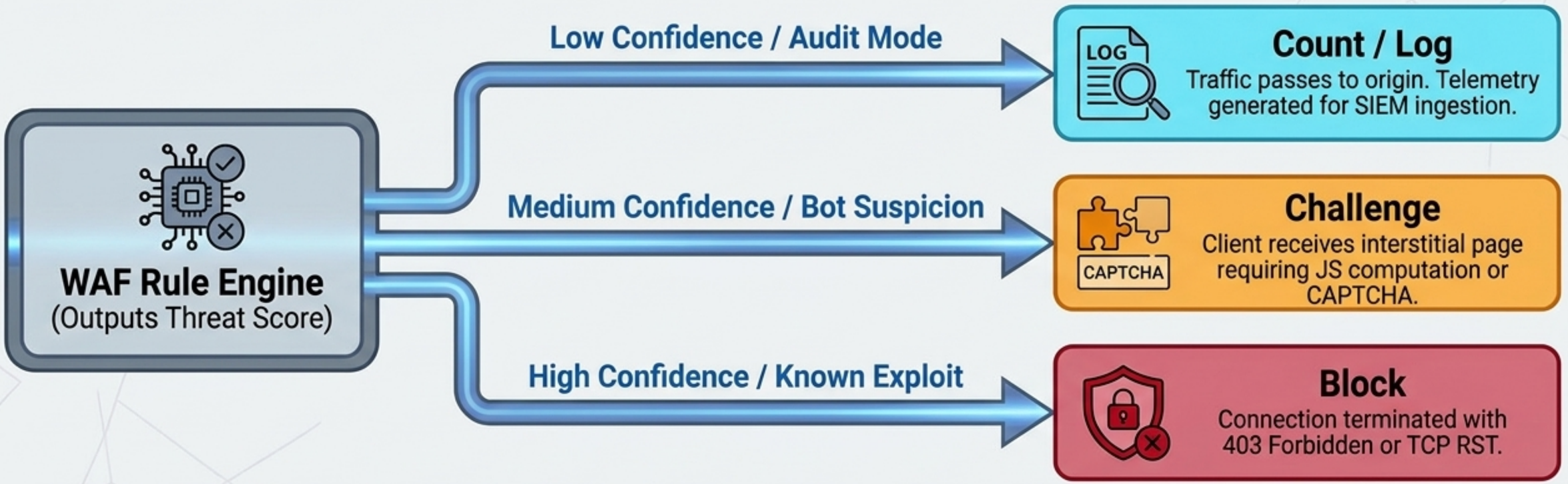
Gemini Notebook

Anti-evasion normalization strips obfuscation prior to rule evaluation



Gemini Notebook

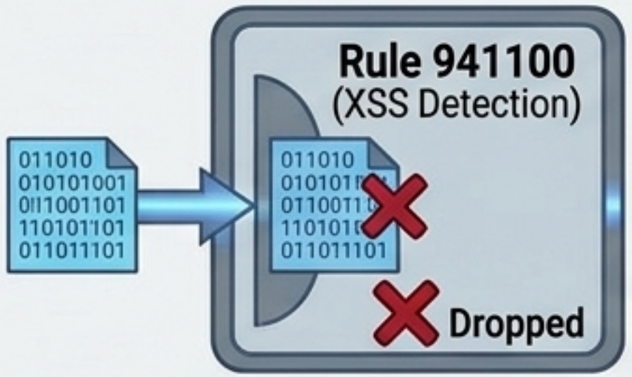
Action routing dynamically dictates responses based on threshold confidence



Negative Security Model: Blocklists denying known attack patterns.
Positive Security Model: Allowlists admitting only pre-approved traffic.

Gemini Notebook

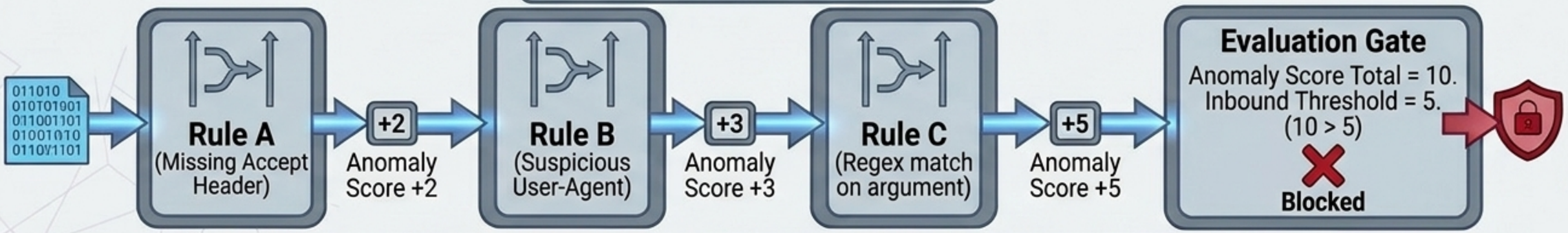
Cumulative anomaly scoring prevents false positives inherent to single-rule blocklists



Legacy Single-Rule Blocklist

High risk of False Positives on complex data.

Cumulative Anomaly Scoring



Minor infractions act as **warning signals**; only aggregate malicious behavior triggers a block, providing balanced **defense-in-depth**.

Gemini Notebook

Paranoia Levels (PL) balance detection sensitivity against operational overhead

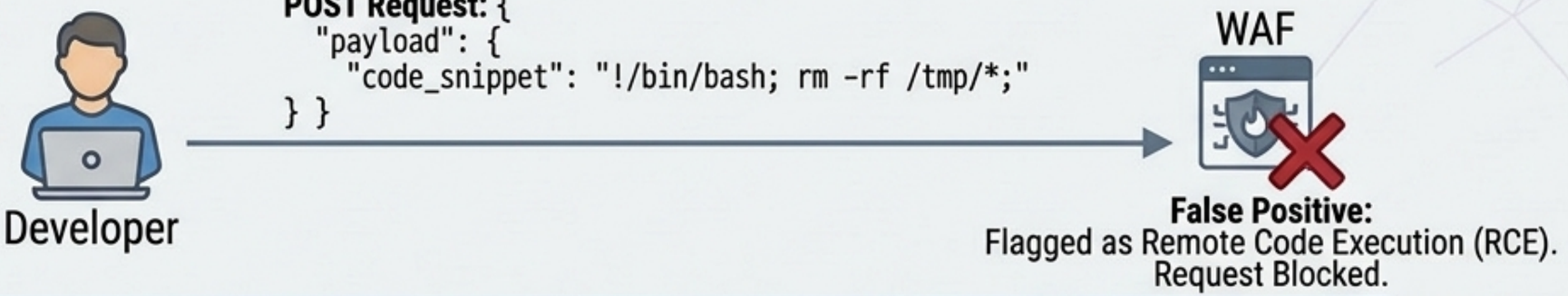
PL1 (Default)	PL2	PL3	PL4 (Max Security)
✓ Base rules. ✓ High reliability, minimal false positives. ✓ Catches obvious, generic attacks. ✓ Target: General web apps.	✓ Adds strict regexes (e.g., restricted headers). ⚠ Catch rate increases, false positives increase. ✓ Target: Standard corporate data.	✓ Tighter limits on special characters/lengths. ⚠ Requires dedicated tuning. ✓ Target: Online banking, healthcare.	✓ Restrictive allowlist-style constraints. ⚠ Guarantees false positives without tuning. ✓ Target: Top secret, isolated enclaves.

Detection Sensitivity & Tuning Complexity Increasing →

Gemini Notebook

Technical payloads require surgical rule exclusions to prevent collateral damage

The Problem



The Conflict

Disabling the global RCE rule to allow this payload creates a massive security vulnerability.

The Fix (Rule Tuning)

Bad Practice:
Disable Rule 932100 entirely.

Good Practice: Surgical Rule Exclusion.

```
SecRuleUpdateTargetById 932100 "!ARGS:payload.code_snippet"
```

The WAF stops inspecting **ONLY** that specific JSON key for **ONLY** that specific rule, **preserving global security**.

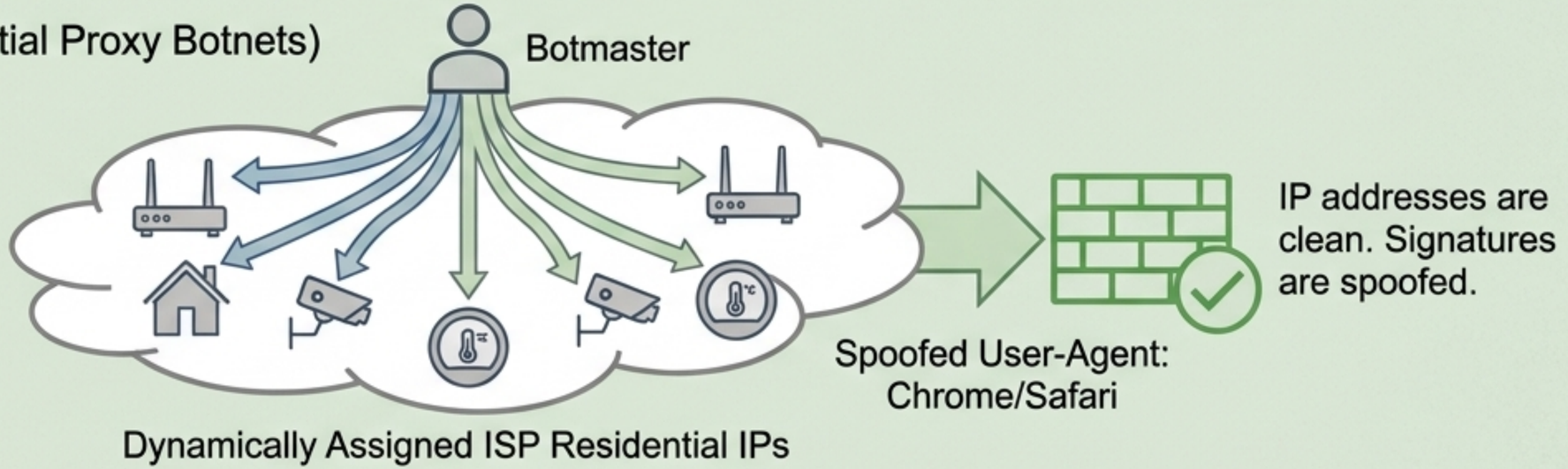
Gemini Notebook

Residential proxies render IP reputation and User-Agent blocking obsolete

Legacy Model (Datacenter Bots)



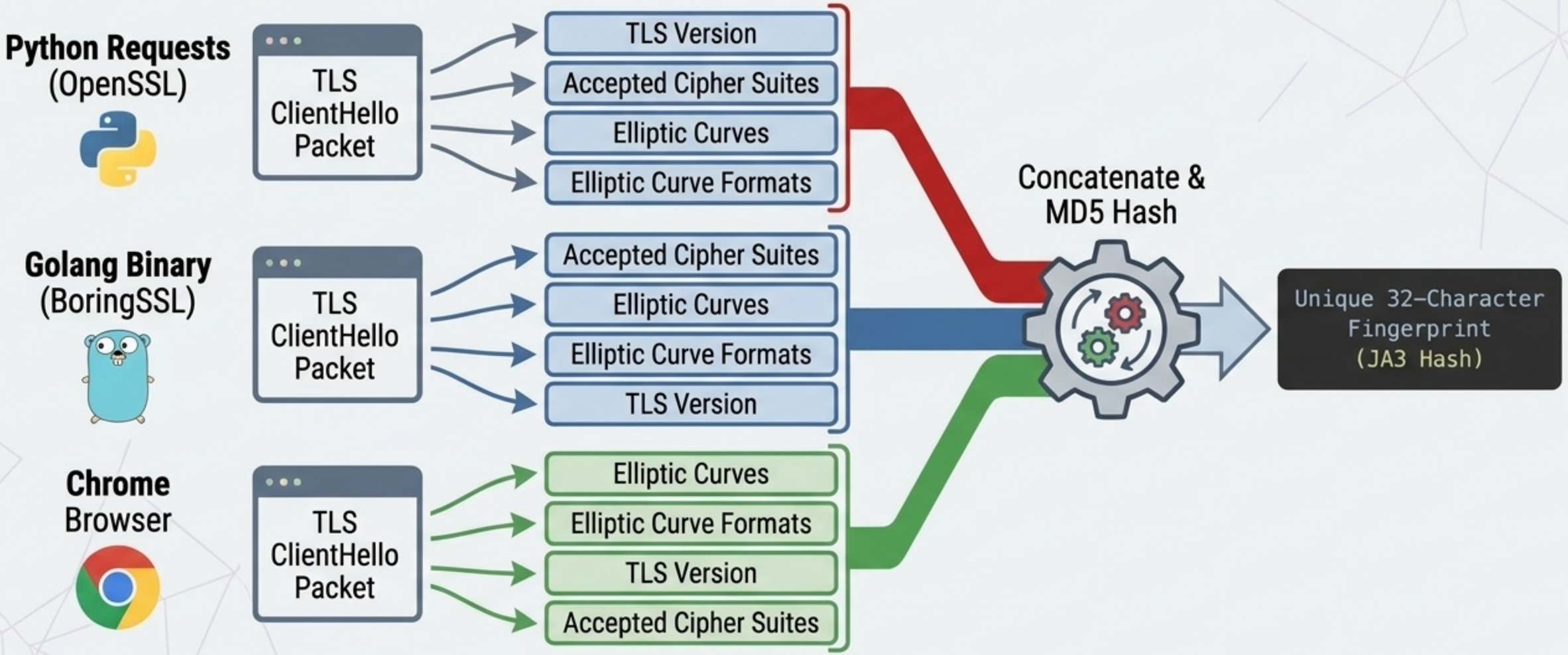
Modern Model (Residential Proxy Botnets)



Rate limiting and blocking by IP address alone causes unacceptable collateral damage (blocking legitimate users on the same ISP) while failing to stop distributed residential botnets.

Gemini Notebook

JA3/JA4 fingerprinting profiles the underlying TLS client library

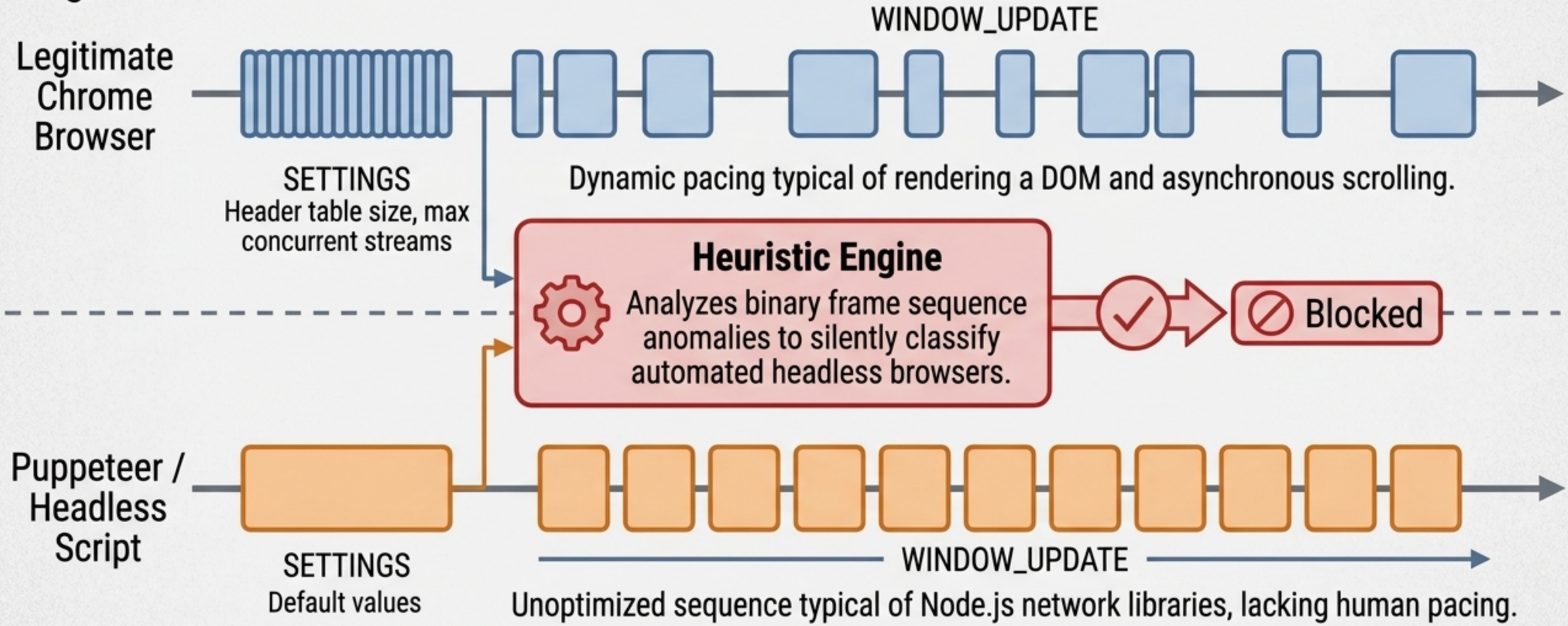


Key Insight: Even if a bot spoofs its User-Agent header as Chrome, the TLS fingerprint will mathematically betray it as a Python script at the network level.

Gemini Notebook

HTTP/2 frame multiplexing exposes automated headless browsers

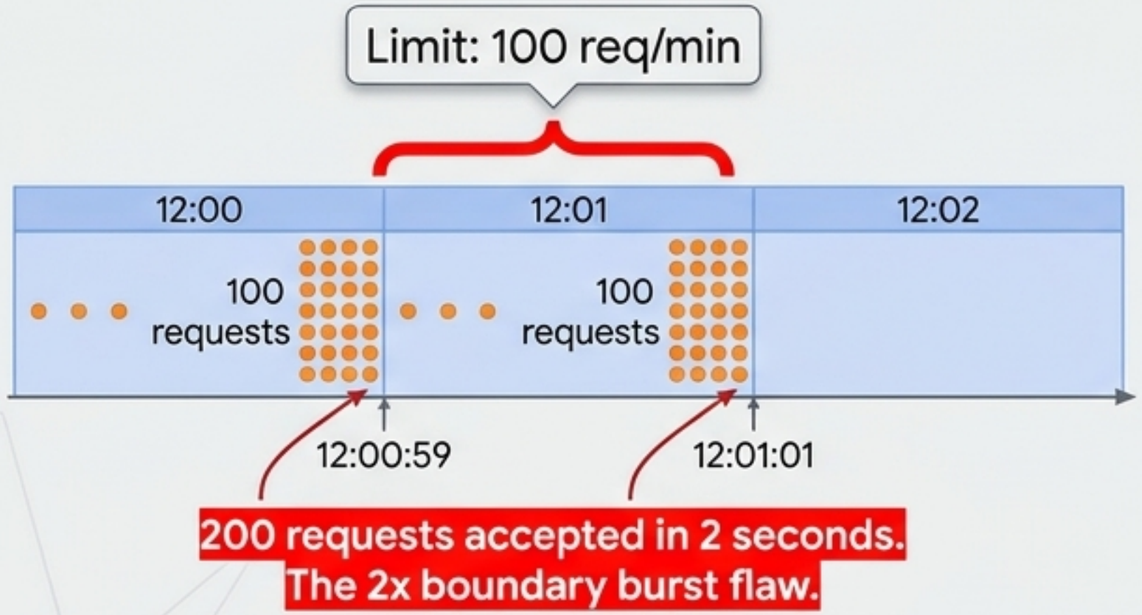
HTTP/2 uses binary frames (HEADERS, DATA, SETTINGS, WINDOW_UPDATE) multiplexed over a single TCP connection.



Gemini Notebook

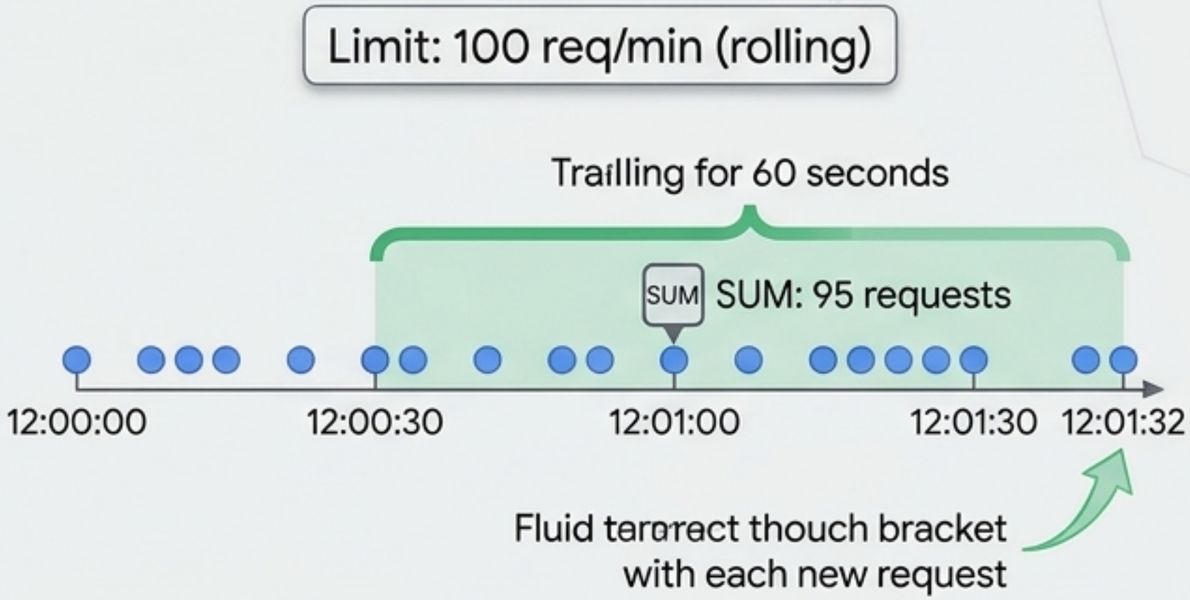
Rate limiting limits traffic, but basic Fixed Windows suffer from boundary burst flaws

Fixed Window Counter

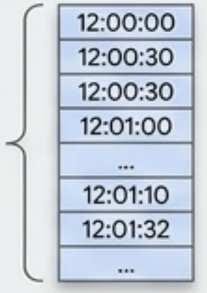


Traffic is constrained within fixed, non-overlapping time windows, leading to a potential doubling of the request limit during window transitions.

Sliding Log



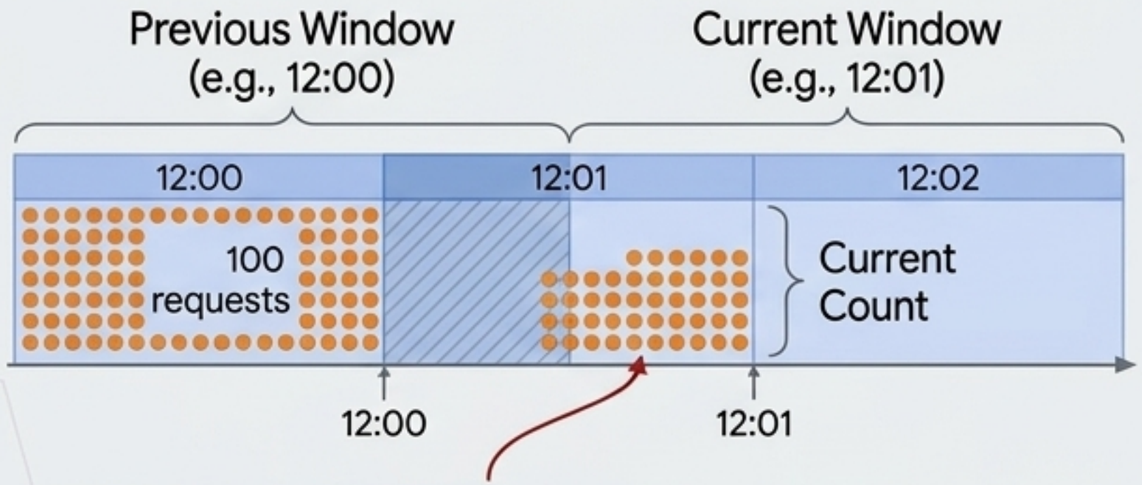
Perfect precision (no boundary bursts), but massively expensive in memory (O(N) storage per user) and compute. Unscalable for high-volume edge networks.



Gemini Notebook

Sliding Window Counters and Token Buckets optimize memory and control burstiness

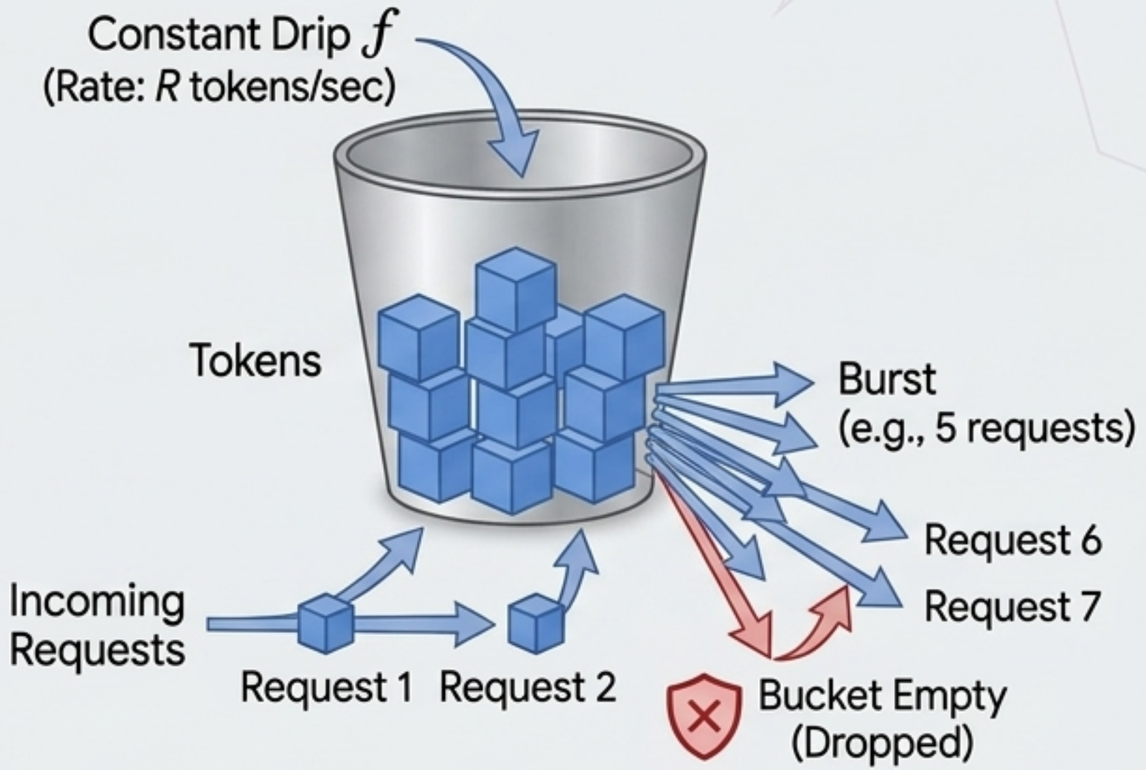
Sliding Window Counter (Approximation)



Total Count = Current Count +
(Previous Count × Overlap Percentage)

→ Blends low memory of Fixed Window with smoothness of Sliding Log. Highly scalable, O(1) memory approximation.

Token Bucket

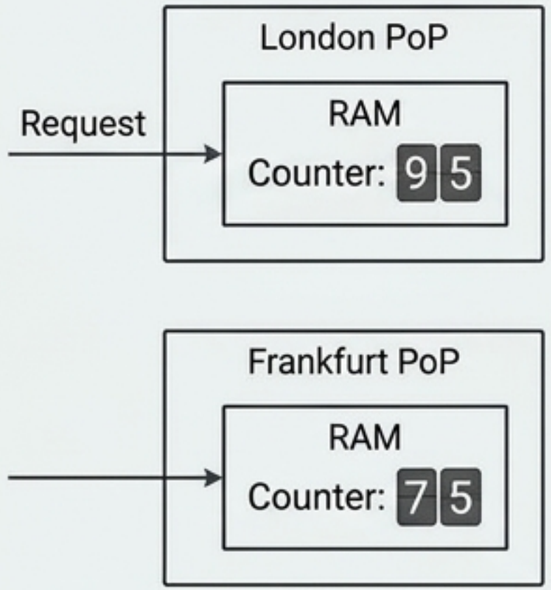


Fluid burst-allowance. Permits short bursts up to capacity, but strictly enforces sustained average rate.

Gemini Notebook

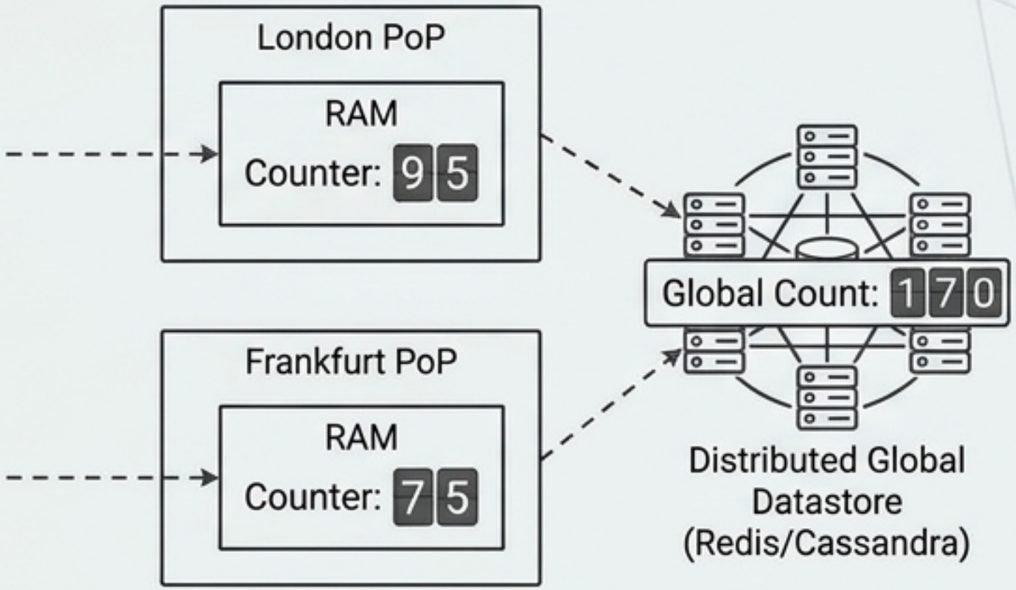
Global state synchronization introduces latency trade-offs for rate enforcement

Local State Tracking



Zero latency, ultra-fast.
Flaw: Attackers can bypass limits by rotating IPs to hit different global PoPs simultaneously.

Global State Syncing



Edge servers asynchronously write counter data to the global datastore.
Trade-off: Introduces CAP theorem challenges. Replication latency (eventual consistency) causes brief windows where limits are exceeded.

Gemini Notebook

Composite key design prevents CGNAT and corporate VPN collateral damage

The Anti-Pattern (IP-Only Limiting)



Blocks the entire company because aggregate traffic exceeds the single-IP threshold.

The Solution (Composite Keys)

Cryptographic Hash Function
Key = Hash(IP Address + Session Cookie + API Key)

Composite Key Mapping



Counters track individual user sessions within the VPN IP. The abusive user is blocked while colleagues continue working without disruption.

Gemini Notebook

Direct-to-origin bypasses negate edge security without cryptographic enforcement

The Vulnerability



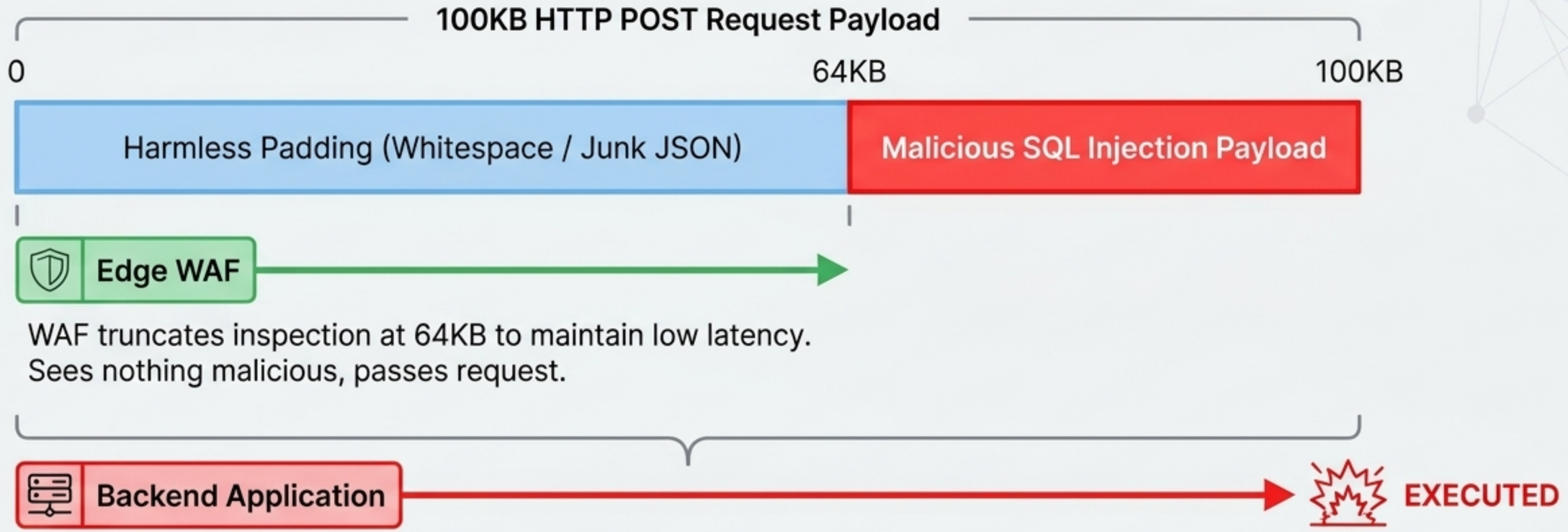
The Fix



- **Network Layer:** Origin Security Groups locked down to accept ONLY WAF provider CIDR blocks.
- **Application Layer:** Authenticated Origin Pulls (mTLS). Origin requires a unique cryptographic client certificate signed by the WAF. Direct requests fail TLS handshake.

Gemini Notebook

Request body truncation and parser differentials create blind spots

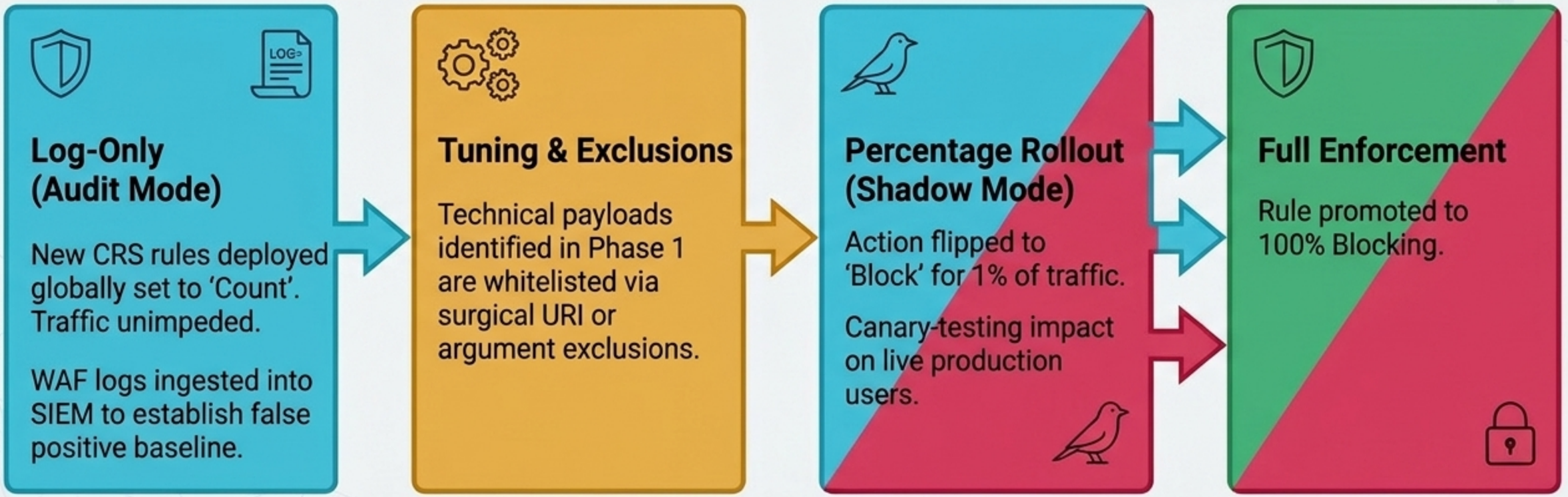


Mitigation: Enforce strict Content-Length limits or configure rules to block if the payload exceeds the WAF's maximum inspection buffer.

Gemini Notebook

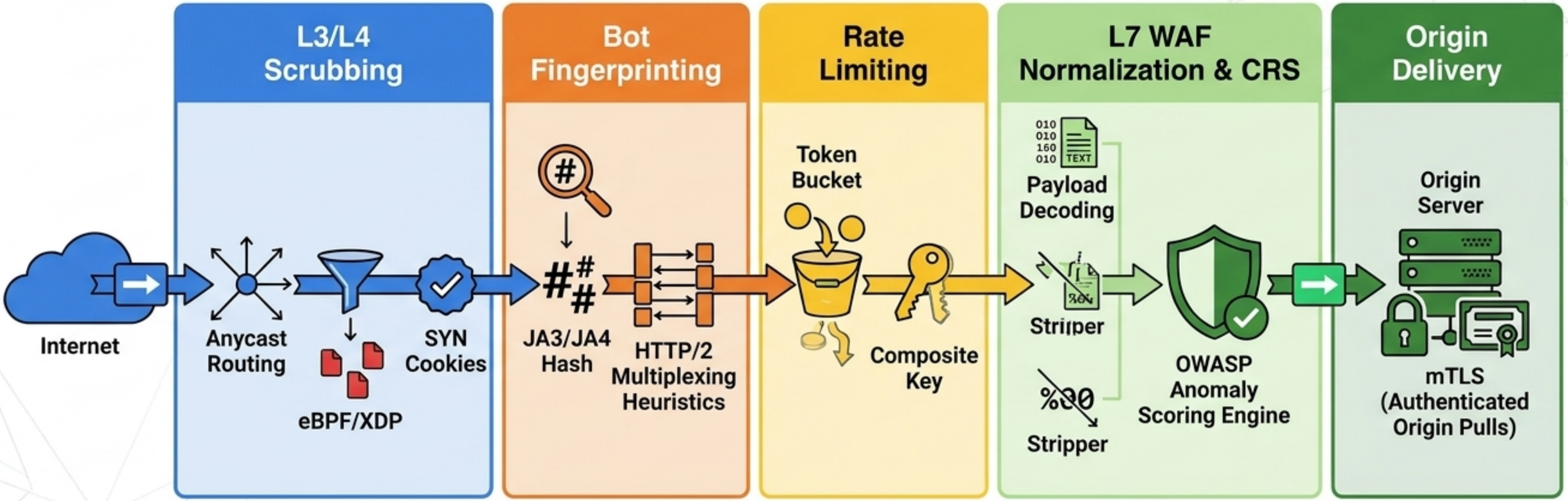
Safe rollout playbooks rely on shadowed telemetry before enforcement

A four-stage horizontal deployment pipeline (Count-to-Block).



Gemini Notebook

The Unified Edge Defense Pipeline




```
> Resilience is not a single perimeter, but a
continuous gauntlet of algorithmic attrition.
```

Gemini Notebook