Apex Systems

# Nexa Cloud Gateway

## Next-Generation API Edge Proxy Architecture

**Presenter:** Dr. Sarah Jenkins, Chief Architect
**Date:** August 2026
**Status:** Under Review (v1.2-RC3)

# Project Overview: Nexa Cloud Gateway

## Context

- Apex Systems is upgrading its edge infrastructure to support high-throughput cloud API routing.
- The Nexa Cloud Gateway acts as the unified entry point for all external client traffic.
- Existing reverse proxies suffer from latency spikes during peak load events.

## Core Mandate

- Consolidate authentication, routing, and rate-limiting into a single high-performance edge layer.

### Key Design Targets

- **Low Latency:** $< 5$ ms p99 overhead.
- **Zero Trust:** Inline token verification.
- **Scale:** $100$k$+$ requests/sec.
- **Resilience:** Multi-region fallback.

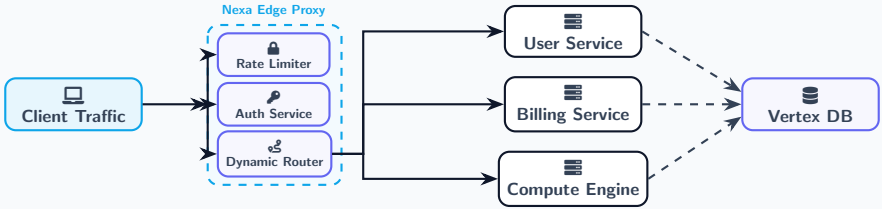
# Key Objectives & Scope

- **Unified Gateway Entry:** Replace the legacy regional proxies with a centralized cluster of Nexa nodes.
- **Dynamic Routing:** Route requests based on header metadata, query parameters, and geographical proximity to Vertex databases.
- **Distributed Rate Limiting:** Prevent service degradation using a global token-bucket algorithm synchronized via Nimbus caches.
- **Observability:** Export real-time metrics, logs, and distributed traces to the centralized monitoring plane.

## Scope Exclusions

This review focuses on the API gateway runtime. Long-term log archiving databases and user billing engines are excluded from this review.

# Nexa Cloud Gateway: High-Level Architecture




- **Unified Inlet:** Traffic is processed at the proxy boundary where auth and rate-limiting filters run concurrently.
- **Downstream Security:** Backend services query the multi-region Vertex DB cluster over isolated virtual networks.

# Protocol & Storage Trade-off Matrix

| Decision Point | Option A | Option B | Core Trade-off | Verdict |
|---|---|---|---|---|
| **API Protocol** | gRPC (HTTP/2 multiplex) | REST (JSON over HTTP/1.1) | gRPC has lower latency but requires client SDK distribution. | **gRPC (Internal)** |
| **Auth Storage** | In-memory Cache (Nimbus Local) | Centralized Cache (Vertex Redis) | Local has sub-ms reads but introduces replication delay. | **Hybrid Layer** |
| **Config Engine** | Etcd (Raft consensus) | Relational DB (Vertex PostgreSQL) | Etcd is highly resilient but adds infrastructure footprint. | **Etcd Cluster** |
| **Rate Limiting** | Token Bucket (Local window) | Sliding Window (Distributed Redis) | Token Bucket is faster; sliding window prevents edge spikes. | **Token Bucket** |

**Architectural Verdict:** Internal inter-service calls use gRPC; external traffic enters via HTTPS/REST. Rate limiting runs locally with asynchronous background synchronization.

# Risk Matrix & Action Plan

| Identified Risk | Severity | Mitigation Strategy | Owner |
|---|---|---|---|
| **Regional Network Partition** Slow database sync across regions. | CRITICAL | Configure Vertex DB for active-passive replication with local read fallbacks. | Infrastructure Team |
| **Auth Token Latency** Inline signature verification overhead. | MEDIUM | Cache verified public keys and use cryptographically signed tokens. | Security Team |
| **Config Deployment Failure** Invalid routing config propagation. | LOW | Implement dry-run verification in CI/CD pipeline before push to Etcd. | Release Ops |

⚠ **Key Security Directive:** No external request shall bypass the rate-limiting filter. Under extreme partition conditions, the gateway fails closed to protect databases.

# Future Work & References

### Key Milestones

- **Q3 2026:** Deployment of multi-region test environment.
- **Q4 2026:** Chaos testing and failure mode verification.
- **Q1 2027:** Full migration of external traffic from legacy gateways.

### Review Action Items

- Design and sign off on rate-limit synchronization interval thresholds.

### References


Meridian, Alex and Lin Zhao (2025). "Scalable API Gateway Architectures for Edge Computing". In: *Journal of Systems Engineering* 14.2, pp. 112–125.

Nimbus, Thomas (2024). *Distributed Edge Systems: Reliability and Security*. Meridian City: Synapse Press.