Nexa Infrastructure Solutions

Capacity Planning & Performance Report

Focus: Vertex Database Cluster & Nimbus API Gateway

Document Ref:	NEXA-2026-CAP-08	**Status:**	Production Draft
Author:	Sarah Jenkins, Principal Architect	**Date:**	August 4, 2026
Reviewer:	Marcus Vance, VP of Engineering	**Version:**	v1.4.2

Abstract

This capacity planning and performance report evaluates the performance metrics and scaling requirements of the Vertex Database Cluster, a core database engine powering Nexa's SaaS platforms. Over the past six months, system workload has grown by 42%, leading to elevated resource utilization during peak traffic windows. This document details the utilization trend analysis, results of automated load testing up to 10,000 requests per second (RPS), and database headroom projections. Finally, we propose a strategic scaling recommendation to migrate the current deployment to high-memory cluster instances, outlining cost projections, performance gains, and implementation risks.

Contents

1 Executive Summary

This capacity planning and performance report assesses the enterprise-grade resource requirements of the **Vertex Database Cluster** and the **Nimbus API Gateway** operated by Nexa Infrastructure Solutions. Over the previous six months (February to July 2026), tenant growth on the Meridian enterprise platform has led to a significant traffic increase. Peak database workloads have surged by 42%, raising serious concerns regarding storage throughput bottlenecks and memory saturation during heavy batch-processing windows.

Key Findings

- **CPU Saturation:** CPU usage during the nightly ETL run regularly exceeds 82% on the primary database coordinator.
- **Storage IOPS Bottlenecks:** Average database write-wait latency has increased from 4.2 ms to 12.8 ms due to disk queue saturation on Standard volumes.
- **Memory Headroom:** Current cache-hit ratio remains healthy at 98.4%, but memory headroom will fall below 10% by November 2026 at current growth rates.
- **Scaling Recommendation:** A transition to High-Memory instance types combined with a migration to Provisioned IOPS SSD storage will resolve performance risks for the next 18 months, representing an estimated net cost increase of 2,400 $ per month.

This report presents detailed historical utilization trends, system architectural details, simulated load testing results, and concrete scaling recommendations. Actionable execution paths are outlined in Section 5.

2 System Overview & Architecture

The system under evaluation consists of the core components of Nexa's cloud infrastructure: the Nimbus API Gateway, the Synapse Caching Layer, and the Vertex Database Cluster. Together, they handle the request-response lifecycle of all Meridian applications. Figure 1 illustrates the layout and flow of data between these components.

2.1 Resource Configuration

The current hardware configurations for the production instances are detailed below:

- **Nimbus API Gateway:** 4× standard VMs (4 vCPUs, 8 GB RAM) configured behind an application load balancer.
- **Synapse Caching Layer:** 3× cache nodes (2 vCPUs, 16 GB RAM) using Redis cluster sharding.
- **Vertex Database Cluster:** 3× high-compute VMs (8 vCPUs, 32 GB RAM, 500 GB SSD storage with 3,000 baseline IOPS).

3 Utilization Trends & Headroom Analysis

We compiled metrics from Nexa's production monitoring suite over the last six months (February 2026 to July 2026) to identify capacity trajectories. A consistent increase in average and peak utilization was observed across the Vertex cluster.

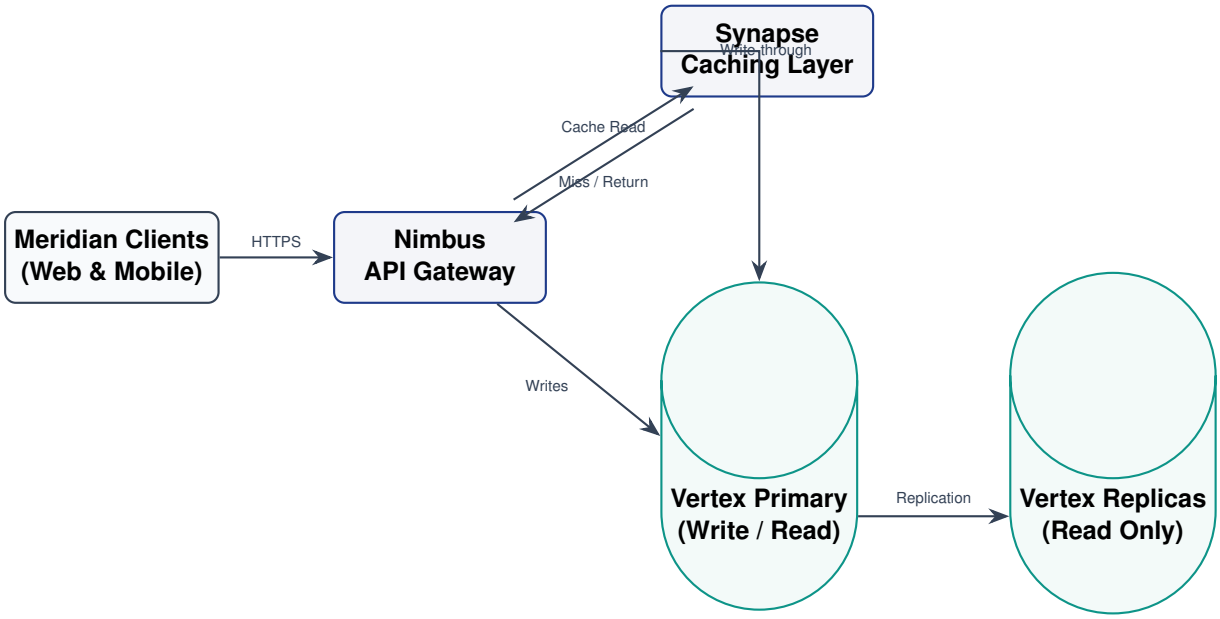

Figure 1: High-level architectural diagram of Nexa's core data-path components.

3.1 Historical Resource Trends

Figure 2 highlights the growth in CPU and memory utilization.

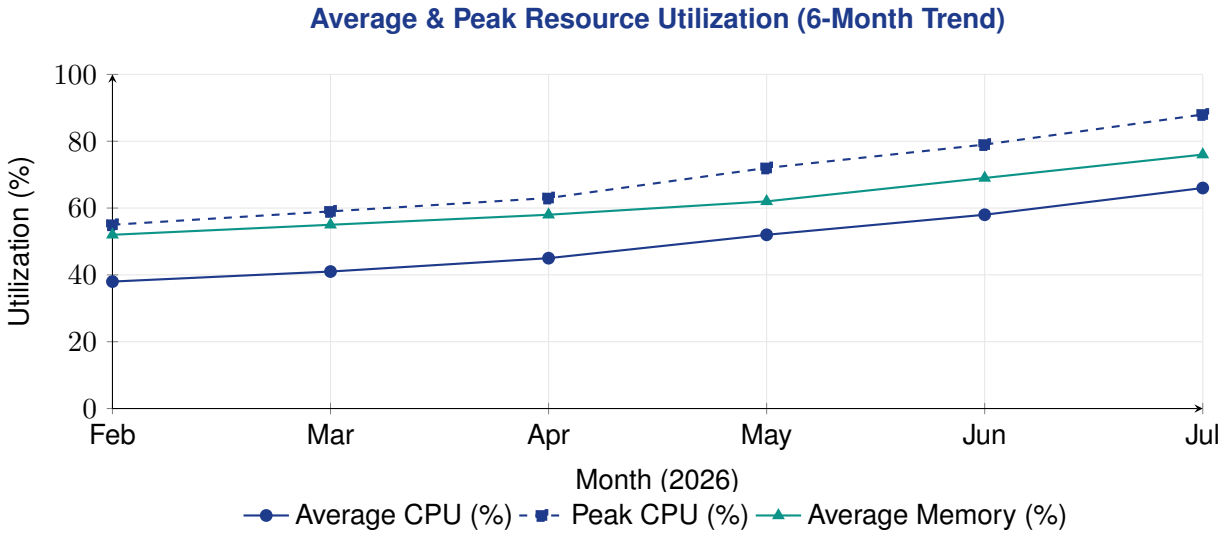

Figure 2: Six-month historical utilization trends for the primary database node in the Vertex Cluster.

3.2 Headroom Analysis

Based on linear extrapolation of our current growth trends (approx. 5% monthly increase in base traffic), we computed the remaining headroom before resource saturation. Saturation is defined as exceeding 85% average utilization or 95% peak utilization.

Table 1 presents the projected saturation dates for the critical system metrics.

Table 1: Vertex Database Cluster resource headroom and saturation projections.

Metric	Current Value	Growth Rate	Provisional Limit	Projected Saturation
Primary CPU (Peak)	88%	+5.1% / Mo	95%	September 2026
Primary Memory	76%	+4.0% / Mo	85%	October 2026
Disk Storage	320 GB	+22 GB / Mo	450 GB (90%)	January 2027
Storage IOPS	2,850	+8.5% / Mo	3,000 (Baseline)	August 2026
Replica CPU (Peak)	46%	+3.2% / Mo	95%	May 2027

Analysis: The disk IOPS metrics and primary CPU peak utilization are critical bottlenecks. High storage write-wait latency has already begun degrading API response times, as detailed in Section 4.

4 Load Testing & Performance Limits

To validate the headroom projections, Nexa's Performance Engineering team performed a series of controlled, synthetic load tests on the staging database environment. The staging cluster replicates the exact CPU and memory specifications of the production Vertex Cluster.

4.1 Methodology

We simulated customer request flows utilizing Apex load-generation agents. The load scenario modeled API read-write requests in an 80:20 ratio. The workload started at 1,000 Requests Per Second (RPS) and was ramped up incrementally over two hours to a peak of 10,000 RPS.

4.2 Latency & Throughput Results

Table 2 summarizes the response latency percentiles and database health status during key milestones.

Table 2: System performance metrics under escalated request loads.

Load Step	Throughput	P50 Latency	P95 Latency	P99 Latency	System Status
Baseline	1,500 RPS	12 ms	45 ms	110 ms	Optimal; healthy queues
Step 1	3,000 RPS	18 ms	62 ms	155 ms	High disk IOPS saturation
Step 2	5,000 RPS	35 ms	188 ms	420 ms	CPU thread contention
Step 3	7,500 RPS	98 ms	750 ms	1,980 ms	Elevated page-faults
Limit	10,000 RPS	450 ms	3,800 ms	8,400 ms	Database connection pool exhausted

4.3 Identified Bottlenecks

The load tests revealed two primary choke points:

1. **Disk I/O Blockages:** Above 3,000 RPS, write operations were stalled waiting for available disk channels. The baseline 3,000 IOPS was exhausted, and storage latencies reached a critical average of 42 ms.
2. **Lock Contention:** At 5,000 RPS, we observed CPU saturation on the Primary coordinator due to row-level locking on the ‘meridian_tenant_balances‘ table. This led to query queuing, pushing the P99 latency past 420 ms.

5 Recommendations & Cost Impact

To mitigate the performance degradation and meet the growth requirements of the Meridian enterprise platform, we recommend a two-phase scaling action plan.

5.1 Proposed Upgrades

1. **Database Node Migration (Compute & Memory):** Upgrade the current Vertex database nodes from 8 vCPUs / 32 GB RAM instances to High-Memory configurations featuring 16 vCPUs and 64 GB RAM. This doubles the compute capability and expands database shared buffers, reducing disk read dependencies.
2. **Storage Tier Elevation (Storage & IOPS):** Migrate database block storage from Standard volumes to Provisioned IOPS volumes. We recommend provisioning 10,000 IOPS per volume to handle peak database writes and eliminate I/O blocking.
3. **Nimbus Rate Limiting:** Configure the Nimbus API Gateway to implement adaptive rate limiting at the tenant level. This protects the Vertex database coordinator from sudden query spikes, maintaining cluster stability.

5.2 Cost-Benefit Analysis

The financial impact of the proposed upgrades is summarized in Table 3. Fictional pricing is based on Nexa’s standard cloud provider agreement.

Table 3: Monthly infrastructure cost comparison for the Vertex DB Cluster.

Component	Configuration	Current Cost	New Cost	Net Monthly Increase
Compute (Primary)	1× vCPU/RAM Instance	360 $	720 $	+360 $
Compute (Replica)	2× vCPU/RAM Replicas	720 $	1,440 $	+720 $
Storage Capacity	3× 500 GB SSD storage	150 $	150 $	0 $
Provisioned IOPS	3× 10,000 IOPS (New)	0 $	1,200 $	+1,200 $
Nimbus Gateway	Rate limiter logic configuration	200 $	320 $	+120 $
Total		**1,430 $**	**3,830 $**	**+2,400 $**

Scaling Timeline

Phase 1 (Immediate): Upgrade the baseline storage to Provisioned IOPS on the Vertex primary node. This resolves the storage write latency problem. (Target: August 15, 2026).
Phase 2 (End of Q3): Perform standard rolling cluster upgrades to migrate node instances to High-Memory formats during the scheduled maintenance window. (Target: September 20, 2026).

References

[1] Amit Patel and Lisa Wong. Distributed load testing and bottleneck analysis on api gateways. In *Proceedings of the Apex Systems Conference*, pages 112–120, 2024.

[2] Robert Smith and Jane Doe. *Enterprise Capacity Planning for High-Throughput Databases*. Meridian Tech Press, San Francisco, CA, 2023.

[3] Marcus Vance and Sarah Jenkins. Performance optimization strategies in cloud-native databases. *Nexa Technical Review*, 14(3):45–58, 2025.