Capacity Planning and Scaling Strategy (Infrastructure)

Department: Operations (SRE Team)
Plan ID: CAP-STRAT-2025Q4
Review Date: Quarterly

1. Core Principles

NexaCore's capacity strategy is built on preemptive scaling to handle the high-volume transactional loads typical of enterprise automation, while mitigating the risk of resource exhaustion or performance degradation.

1. **Predictive Analysis:** Base capacity planning on 12-month trailing client growth data and 6-month product adoption forecasts (e.g., predicted workflow count per client).
2. **Headroom Buffer:** Maintain a minimum of **20%** spare capacity (Headroom) across the cluster before triggering an alert for manual intervention or purchasing Reserved Instances.
3. **Client Tiering:** Capacity for high-tier enterprise clients (Tier 1) is guaranteed and monitored separately from the general pool.

2. Key Metrics for Capacity Monitoring

Metric	Threshold/Alert	Action Required
Cluster CPU Utilization (Avg.)	for 30 minutes	Trigger **Capacity Alert** (P2). Review autoscaling configuration.
Database Connection Pool Usage	for 10 minutes	Trigger **Performance Alert** (P1/P2). Review query efficiency and connection limits.
Disk IOPS/Throughput	of provisioned limit	Trigger **Degradation Alert** (P1). Immediately provision higher IOPS storage.
AI Model Inference Latency	(P95)	Trigger **Performance Alert** (P1). Scale up inference microservice workers.

3. Scaling Mechanisms

Mechanism	Trigger	Response Time	Description
Horizontal Pod Autoscaling (HPA)	CPU/Memory utilization above (Container Level)	Instantaneous (Seconds)	Automatically adds or removes application pods within the cluster.
Cluster Autoscaling (CA)	Cluster node utilization above (Node Level)	Fast (Minutes)	Adds new virtual machine nodes to the Kubernetes cluster to accommodate new pods.
Pre-Scheduled Scaling	Known events (e.g., weekly data ingestion ETL job, major marketing campaign load)	Pre-configured	Manually increase cluster size 1 hour prior to the event and scale down afterwards.
Reserved Instance Purchase	Predictive planning shows base capacity exceeding RI coverage.	Quarterly	Long-term commitment for guaranteed, discounted capacity.

4. Capacity Review Workflow

Frequency	Activity	Outcome	Owner
Daily	Review real-time resource usage dashboards.	Identify short-term anomalies and immediate scaling needs.	On-Call SRE
Weekly	Review weekly usage trends and cost consumption.	Adjust autoscaling parameters and identify	Ops Lead

		underutilized resources.	
Quarterly	Formal Capacity Planning Meeting (Ops, R&D, Product).	Update the annual budget and 6-month Reserved Instance plan.	Head of Operations